

Interactive Video Mirrors for Sports Training

Perttu Hämäläinen

Helsinki University of Technology

P.O.Box 5400, FIN-02015 HUT

perttu.hamalainen@hut.fi

+358 50 596 7735

ABSTRACT

This paper studies gesture and speech controlled video for sports training. The goal is to combine the benefits of recording your performance with video equipment and training with a mirror. For example, a delayed camera view projected on a screen can be used to repeatedly perform and evaluate a spin kick, a move that is difficult to practice with a mirror.

A video mirror can also be augmented with speech or gesture control for playback, recording and inspecting of individual frames. Three different interface design approaches are evaluated, based on testing with eight users that practice martial arts and acrobatics. The results suggest that an interactive video mirror can be highly useful in martial arts and other sports. The paper also introduces new kind of graphical controls that float around the user so that they can be manipulated with gestures regardless of the user's position.

Author Keywords

Martial arts, speech and gesture control, motion analysis.

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User Interfaces--input devices and strategies, interaction styles.

INTRODUCTION

Mirrors and video analysis are used in many sports to spot errors in pose and motion. In research literature, several computer assisted motion and biomechanics analysis systems are described. Various approaches include user-assisted video analysis, tracking devices, and computer vision [5,7,9,12]. There is also commercial software for analyzing sport videos [3,4]. However, all the systems operate off-line so that you record a video and then analyze it on a computer with a traditional user interface.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

NordiCHI '04, October 23-27, 2004 Tampere, Finland
Copyright 2004 ACM 1-58113-857-1/04/10... \$5.00

This paper shows how interactive video mirrors can combine the benefits of mirrors and video in repeated performing and evaluation of acrobatic and martial arts moves. By a video mirror we mean a setup like the one in Figure 1, where the camera view is shown on a screen in front of the user. If the camera is placed on the side, you can see your profile while looking forward, which helps in martial arts where you should focus on the direction of your opponent. The video output can also be delayed so that you can perform a spin kick and immediately view the move after finishing it. Using gesture and/or speech control, playback and recording capabilities can be added without the athlete having to leave the training area to operate a computer or video equipment.

In the following, three different interface design approaches are compared, including speech commands, gesture operated graphical controls and a motion activated automatic system. The developed systems aim at optimizing the feedback cycle and learning process. In the future, the presented ideas will be integrated in a computer vision based martial arts game installation to motivate learning through gaming. The game explores film-style exaggerated movements in a profile view with several displays. Images and video of the prototype can be found at www.kickasskungfu.net.



Figure 1. Test setup, showing camera (1), camera view projected on a screen (2), user (3) and test instructor (4).

In general, there has been much research in perceptive user interfaces and speech-based interaction [6,10]. Interactive video 'magic' mirrors have been used in games and art installations starting from VideoPlace by Krueger et al. [8], where the two-dimensional video image of the user interacts with computer generated animated characters. The approach has been commercialized by several companies, the latest incarnation being the Eye-Toy camera and games for the PlayStation 2 game console [2]. The MIT Media Lab Alive system is an example of a more sophisticated, 3D interactive mirror where you can interact with computer generated characters using gestures [13].

TEST SYSTEM

The test setup is shown in Figure 1. The software runs on a 2.8GHz Pentium 4 Windows laptop with an USB webcam operating at 30 frames per second.

Motion activated automatic recording and playback

The first system developed was a simple non-interactive mirror with a delay, similar to the Ideo dressing room mirrors, where the delayed mirror image allows customers to see their back when turning around [1]. This was pre-tested by the author and two users and found useful, but rather limited. It would often be useful to view a move several times, optionally in slow-motion.

A simple system was developed that estimates the amount of motion, computed as the sum of absolute differences of pixel intensities in two most recent video frames. If the motion exceeds a threshold, the system starts recording. When there is no motion for a specific time interval, recording stops and the system starts to play the recorded video in a loop, showing every other replay in slow-motion.

Spoken commands

The motion activated automatic system was pre-tested with two users and mixed results were obtained. The system seemed promising, but it also sometimes reacted to unintentional motions. To make the system more robust, alternative interfaces with explicit commands were designed. For unobtrusive control from the training area, there are basically two alternative approaches: gestures and speech.

Speech recognition was implemented using the Microsoft Speech API. The available commands are 'rec', 'play', 'stop' and 'delay', corresponding to the possible states of the system. In playback mode, the recorded video plays in a loop with every other replay in slow-motion, similar to the automatic system.

Gesture control: floating overlay widgets

Eye-Toy menus use robust and intuitive motion activated buttons overlaid in the camera view, which acted as a starting point for the gesture based interface. In addition to the basic static buttons, the interface has an optional dynamic mode where the buttons move with the user, staying at a constant distance. The framing of the camera

view is looser than in Eye-Toy to give you space to perform large motions, so static buttons cannot be reached from all locations. Compared to previous systems, this is also the first time overlay controls are evaluated in a profile view so that if the control is on your right on the screen, you actually have to reach in front of you.

The interface has four buttons corresponding to the four spoken commands of the speech interface. The screen is split into live view augmented with the buttons and a separate output view showing the recorded or delayed video. The interface also features a vertical slider for finding and inspecting individual frames of recorded video, as shown in Figure 2. The slider is only visible in playback mode so that it does not interfere with training.

When the software is started, a sample image of the background is stored in memory and for each frame, a pixelwise absolute difference to the background is computed. The difference image is thresholded to obtain a binary image where each nonzero pixel is considered to be part of the user. Note that this is only the most basic form of background subtraction, of which a good review is given by Toyama et al. [11].

The buttons are activated if there are enough user pixels inside them. The slider handle is positioned at the mass center of the user pixels inside the slider area. Accuracy is improved and noise reduced by filtering the obtained position with a first order IIR low-pass filter,

$$y'(n) = 0.9y'(n-1) + 0.1y(n), \quad (1)$$

where $y'(n)$ is the filter output at video frame n and $y(n)$ is the y coordinate of the mass center.

In static mode, the gesture widgets are positioned at the top of the image. In dynamic mode, the median height of the bounding box of the user is computed and the widgets are placed at a constant height relative to it, centered at the

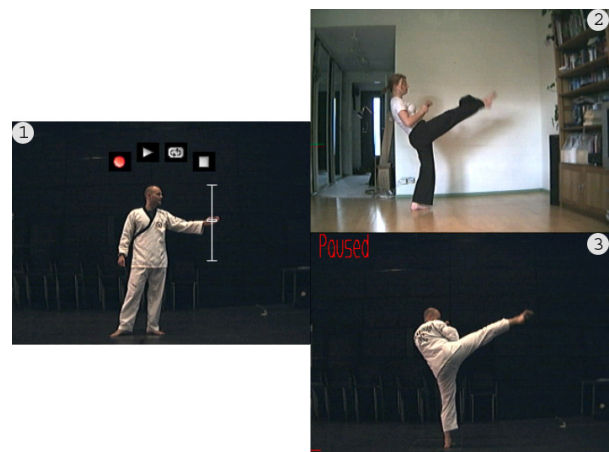


Figure 2. Inspecting individual frames with a gesture operated slider. The display is split into the camera view with overlay widgets (1), model view (2) and inspected frame (3).

user's mass center. The median corresponds to head level if the user's hands or feet are not held above head for most of the time.

Optional model view

The role of a model performance was also inspected. All the three interfaces were augmented with an optional model view, as shown in Figure 2. The obvious disadvantage is that the size of the windows becomes smaller than in speech and automatic modes where there's normally only one window. The model view shows a selected move playing in a loop.

If the model is used in delay mode, the delay is set equal to the length of the model loop so that if you perform a move in sync with the model, you can also watch the delayed video in sync.

TEST PROCEDURE

The system was tested with eight users with three to twenty years of experience in martial arts. Two users also practiced cartwheels and other acrobatic moves. Three users had black belts: 5 dan in karate, 2 dan in jujutsu and 1 dan in taekwondo. None of the subjects had used speech-based interfaces before. Two of the subjects had used gesture interfaces (the Eye-Toy). Both qualitative and quantitative data was collected.

For each user, the three interfaces were shown in random order. The intuitiveness of the interfaces was evaluated so that the users were not given any instructions but to think aloud, try out the system and tell the test instructor how it works. However, in the speech operated mode, the speech commands were given, since they are impossible to communicate without text or sound. The background information given was that the users were evaluating a computerized video system for training. The users were asked to perform both straight and spinning moves where they have to take their eyes off the screen.

The delay mode was compared to a live view for both straight and spinning moves. Different delays were tried in to find the optimal value for each user.

After trying out all the three systems, the users were interviewed. The systems were ranked and their good and bad sides evaluated. The systems were then tried again briefly with the model view added. The role and usefulness of the model view was then discussed.

During the first tests, the speech recognition technology was found unreliable. Although it worked fairly well if you spoke close to the low-cost multimedia microphone used, the recognition was completely random when commands were spoken from a couple of meters away, even though the space was a quiet and acoustically damped studio. Eventually, the tests were conducted in Wizard of Oz manner so that the instructor controlled the system with a keyboard according to the spoken commands.

RESULTS

The ranks given to the systems are shown in Table 1. According to mean rank, the speech-based system performed best and the automatic system was the worst. Speech control was ranked best four times and gesture control three times. A clear advantage of speech commands is that the commands can be given from any location and pose. Speech commands are also close to how you normally interact with a person using a video camera.

Gesture	1	3	1	1	2	2	3	3	2
Speech	2	1	2	2	3	1	1	1	1.6
Automatic	3	2	3	3	1	3	2	2	2.4

Table 1. The ranks given to the three compared systems. The last column contains mean ranks. Smaller rank is better.

It should be noted that the numbers are only suggestive because of the small sample size. The compared systems were only examples of possible speech or gesture user interfaces and the ranks should be considered together with other interview data and observations.

All subjects understood the automatic system without explaining. However, there were problems when a move ended in a different position than it started from. The subjects often stopped for a while and started watching the replay, but then moved back to starting position, which switched the system back to record mode prematurely.

Basic gesture and pose recognition was originally ruled out because of the wide and unknown variety of expected moves and stances in martial arts, let alone sports in general. Ceasing of movement was identified as the only robust cue that could be used so signal a state change. However, the test showed that at least in martial arts, a basic relaxed and unguarded standing position is actually one highly unlikely pose, except when watching the replays. This could be exploited to make the automatic mode more robust. On the other hand, a karate instructor with 20 years of experience ranked the automatic mode as the best one. He considered the system fairly easy to use and remarked that if you perform a move correctly, you should feel comfortable watching the replay from the ending position.

Six subjects understood the overlay buttons without instructing. It was remarked that the buttons are more intuitive than speech, since they visualize the available options and the state of the system. One user also felt that spoken commands were 'from another world' than the otherwise physical and visual training and therefore it was more natural to use the system with gestures.

Six users preferred the moving overlay buttons instead of static ones. The motion of the buttons seems to signal that you can use them. The subjects that preferred the buttons to be static had problems reaching for them. They also touched the buttons unintentionally, e.g., with high kicks

and cartwheels. One subject noted that everywhere he reaches with his hands, he can also reach with his feet. In a future system, the static buttons should be provided as an option and a delay should be added so that the buttons are not activated by a fist or foot swishing through them.

All users first tried to reach for the buttons in screen coordinates. It was most difficult to reach buttons that are both above you and to your right or left, since pointing your hand outside the camera plane decreases reach. After instructing, all users got accustomed to the system, although there were differences in the time needed. The use of gesture widgets in profile and other unconventional views is a topic for further research.

The slider produced spontaneous comments, such as “*this is really fun*” and “*quite an enjoyable device*”. Continuous controllers are also the strongest advantage of gestures over speech.

The delayed camera view was found particularly useful for repetitive training because there is no waiting or worrying about the state of the system. All subjects found a constant amount of delay they preferred for all moves. Delays between 1 and 2 seconds were preferred with a mean of 1.5 seconds. Half of the users felt that they could concentrate better with a delayed view than a live view even when practicing straight kicks.

All users preferred to have the model video visible at all times rather than watch a model video and then practice focusing on their own performance. Although live and delay views with a model were found useful when starting to practice a move, it was difficult to compare the two displays. For comparison, slow motion and frame-by-frame playback were more useful.

CONCLUSION

Three alternative designs of an interactive video mirror for martial arts and acrobatics were presented and evaluated. In general, the concept was found highly useful. All test subjects also considered the system useful for dance and other sports where you must polish up moves and postures.

The results suggest that the floating overlay widgets introduced in this paper are a promising approach, although the target environment must be considered carefully. For example, one user criticized the overlay widgets for the smaller size you see yourself in because of the split screen display. This can be a problem, particularly if a monitor is used instead of a projected display. On the other hand, a split screen display is required anyway with a model view.

Speech was ranked best by half of the subjects, but the limitations of speech recognition technology must be considered in practice. With a limited set of commands, the current systems work in general, but noise from the environment and consumer grade hardware can still cause problems. Training the recognition engine helps, but managing of multiple user profiles becomes a problem for

example in schools and public installations. Speech recognition is also robust only when training alone and not speaking to your fellow trainers.

Two users suggested a pedal controller as a future improvement. However, accidentally stepping on a pedal can cause injury, for example, when practicing somersaults. A wearable or wrist-held wireless remote controller could also be an option, but perceptive technology is more flexible with several users taking turns. Clapping your hands was also suggested as a control method.

ACKNOWLEDGEMENTS

This work has been supported by HeCSE graduate school.

REFERENCES

1. Ideo Prada RFID Closet, http://www.ideo.com/case_studies/prada.asp?x=5
2. Sony Eye-Toy, <http://www.eyetoy.com>
3. Sports Motion Inc., <http://www.sportsmotion.com>
4. WINalyze, <http://www.winalyze.com>
5. Baecker, R., Miller, D., and Reeves, W. Towards a Laboratory Instrument for Motion Analysis. *Computer Graphics*, vol. 15, No. 3, August 1981, pp. 191-197
6. Crowley, J.L., Coutaz, J., Bérard, F., Things That See, *Communications of the ACM*, Vol. 43, No. 3, March 2000
7. Iredale, F., Farrington, T., and Jaques, M., Global, fine and hidden sports data: applications of 3-D vision analysis and a specialised data glove for an athlete biomechanical analysis system, *Proc. Mechatronics and Machine Vision in Practice* (1997), pp. 260-264
8. Krueger, M., Gionfriddo, T., Hinrichsen, K. VIDEOPLACE - An Artificial Reality, *Proc. CHI'85*, ACM Press (1985), pp. 35-40
9. Perš, J., Bon, M., Kovačič, S., Šibila, M., Delman, B. Observation and Analysis of Large-scale Human Motion, *Human Movement Science*, 21(2), 295-311, July 2002
10. Rosenfeld, R., Olsen, D. and Rudnicky, A. Universal speech interfaces, *Interactions*, Vol 8, Issue 6, October 2001
11. Toyama, K., Krumm, J., Brumitt, B., and Meyers, B. Wallflower: Principles and Practice of Background Maintenance, *Proc. International Conference on Computer Vision* (1999)
12. Yamamoto, M., Kondo, T., Yamagiwa, T., Yamanaka, K. Skill recognition, *Proc. Automatic Face and Gesture Recognition* (1998), pp. 604 – 609
13. Wren, C.R., Spacarino, F., Azarbayejani, A., Darrel, T., Davis, J., Starner, Kotani, A., Chao, C., Hlavac, M., Russel, K., Bobick, A., Pentland, A., Perceptive Spaces for Performance and Entertainment (Revised), *Applied Artificial Intelligence*, Vol. 11, No. 4, June 1997