A Study of Visual Context Representation and Control for Remote Sport Learning Tasks

Wanmin Wu, Zhenyu Yang, Klara Nahrstedt
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois, USA
{wwu23, zyang2, klara}@uiuc.edu

Abstract:

Shared visual context plays a vital role in video-mediated remote learning tasks. However, it remains unclear *how* to provide such visual context in order to best facilitate learning. In this paper, we present an exploratory study to (a) evaluate the impact of visual context *representation* on sport learning in video-mediated environments, and (b) investigate how users *control* the visual context (e.g., changing viewpoint) in assisting learning. We find that (a) for sports, the 3D free view representation enables faster, more accurate learning than 2D representation; (b) average users tend to "browse" randomly instead of selecting a particular viewpoint with intention; and (c) different users have dramatically different control patterns for similar visual context, which suggests the importance of user customization. We anticipate our results will bring theoretical and practical implications for the design of next-generation video-mediated learning systems.

1 Introduction

Technology is bringing people closer than ever. A group of geographically distributed students can sit "together" in a virtual classroom for a lecture given by remote instructors [Microsoft Conference XP; Chen 2001]. A young man can learn Tai-Chi in a 3D tele-immersive environment with a professional coach overseas [Jung et al. 2006]. These are made possible by advanced video-mediated systems that provide "shared visual context" for users to collaborate or interact. Prior empirical literature has emphasized the strong value of shared visual context in a range of remote collaborative tasks (e.g., [Fussell et al. 2000; Kraut et al. 2003]). However, it remains unclear *how* to present the visual context to best facilitate learning of physical activities.

Conventional video-conferencing systems rely on 2D visual context. Each site is captured by one or more 2D cameras, and all 2D video streams construct a more comprehensive representation of all participating sites (e.g., [Chen 2001, Yu 2006]). More recently, tele-immersive technology has emerged as a new video medium which can provide possibly richer visual context in 3D format [Towles et al. 2002; Yang et al. 2006]. To realize this, an array of 3D cameras is set up in each participating site to capture the scene in 3D from a wide field of view. Furthermore, the video streams from all sites are aggregated and integrated into a shared virtual environment to provide users a sense of shared presence. More generally, we characterize the format of visual context representation into three main categories based on the camera conditions in each participating site:

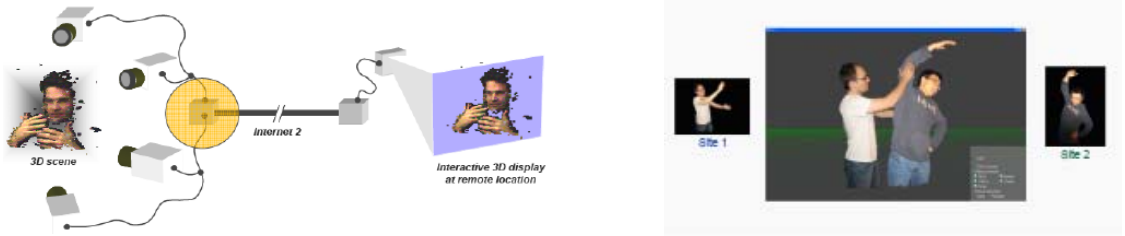
- *Fixed View*: Most systems install a single non-PTZ 2D camera at each site, thereby generating a fixed

view for the aggregated visual context. Virtual Auditorium [Chen 2001], Digital Amphitheater [Mankin et al. 2000], and Breakout for Two [Mueller et al. 2007] are examples of such systems. Only a single stream is presented to the remote participant with a fixed viewpoint. Fixed view stimulates situation awareness by giving high level overview of the scene, but is constrained by the position and orientation of the camera.

- *Multi-2D View*: To address the limitation of fixed view, some systems install multiple 2D cameras to capture a scene from more angles [Gaver et al. 1993; Ranjan et al. 2006; Yu 2006]. Thus, multiple video streams of one scene can be presented to remote users in separate windows. Such multi-2D view increases situation awareness and reduces occlusion of objects by providing more perspectives of a remote scene. However, it may either require user input to switch between cameras [Ranjan et al. 2006] or endure the risk of being visually too complicated if most camera views are shown simultaneously [Chen 2001, Mankin et al. 2000]. In addition, the representation of the visual context is still constrained by the physical placement of the 2D cameras.
- *3D View*: To overcome the constraints of fixed view and multi-2D view, some recent systems seek a different approach by reconstructing a 3D model of each participant, thereby offering a *free-viewpoint* 3D representation of the integrated scene [Baker et al. 2005, Yang et al. 2006]. These systems, called "tele-immersive systems", render all participants into a shared 3D graphical space (*cyber-space*) so that they can interact "virtually" (see Figure 1(a)). The users are allowed to manipulate the view of the cyber-space arbitrarily and hence observe the scene from almost any angle.

Although these three formats of visual representation have been used extensively in many video-mediated systems, we have not found any previous work that compared them for remote learning tasks. It is unclear if the visual representation format has any impact on the performance of learning. In this paper, we fill in the gap with an exploratory study in remote basketball training/learning. Though our ultimate goal is to automate camera control and view selection in a video-mediated system, we believe that the results will be important for a better understanding of the visual representation in video-mediated learning.

2 Background and Related Work



(a) Tele-immersive system (b) An illustration of two-site tele-immersion

Figure 1: 3D Tele-Immersive Systems

Figure 1(a) illustrates a typical 3D tele-immersive (3DTI) system with three main components: (a) capturing, (b) transmission, and (c) rendering. An array of cameras is set up to shoot the scene synchronously in 3D (the *capturing* component). The 3D video streams from all cameras are then aggregated and transmitted to a remote site through Internet2 (the *transmission* component), and finally get rendered on the displays (the *rendering* component). All three components in the system work in high synchronization and coordination in real time to enable the remote interactivity.

Fussell *et al.* [Fussell et al. 2003] identified two ways that visual information facilitates collaboration. First,

it provides *situation awareness* - an ongoing awareness of the work environment and the activities taking place within it. Second, it helps *conversational grounding* - the interactive process by which communicators come to a state of mutual understanding. While visual context is naturally shared in face-to-face co-present interaction, it remains unclear how to provide the context in video-mediated collaborative environments. We believe that the "optimal" visual representation depends largely on the nature of the collaborative learning tasks. For example, face-to-face views may be sufficient for video-mediated negotiation [Veinott et al. 1999], but not for collaborative internal design [Gaver et al. 1993]. Previous studies focused on constructional learning tasks which involved identification and manipulation of small objects such as puzzles and Legos (e.g., [Ranjan et al. 2006]). In this work, we investigate a very different learning task - learning of sport activities, which involves lots of full-body physical movement. In addition, the previous studies focused on using head-mounted cameras and bird-eye cameras to give detailed view and high-level overview of the scene, respectively, which only involved fixed view and multi-2D view by our definition. With the emergence of 3D tele-immersion, we would like to also compare the 2D representation with its possibly richer 3D counterpart.

3 The Current Study

3.1 Methodology

Our goals of the study are twofold: (a) evaluate the impact of visual context representation on sport learning in video-mediated environments, and (b) investigate how users control the visual context (e.g., changing viewpoint) in assisting their learning. More specifically, we aim to understand how different visual context representations affect the performance of sport learning. Furthermore, we want to analyze behaviors of users, particularly how they manipulate the visual context, and examine their behavioral patterns.

We deployed a video-mediated system which could run in three modes, providing fixed view, multi-2D view, and 3D view, respectively. Six pairs of participants (a coach and a student) were recruited from a major university in the United States to take part in the study. Basically, the coach and the student communicated via a shared audio channel and the shared visual context that captures the student's motion in real time. The coach then instructed the student to learn basic basketball skills through the two channels. Three visual representation formats were tested: fixed view, multi-2D view, and 3D view.

For the first goal, we evaluated the learning performance of the students quantitatively and qualitatively. We measured the learning completion time and effectiveness, and used those as the metrics for comparing the different visual context representations. Furthermore, questionnaires were distributed to all participants to collect subjective responses. For the second goal, all the actions made by the users on the shared visual context (i.e., view changes) were logged for post-analysis. We performed extensive analysis on the data to look for behavioral patterns and insights.

3.2 Participants and Procedure

Twelve participants were recruited from a major university in the United States to take part in the study. They were divided into six pairs, where the person with little experience of basketball (< 1 year) became the student, and the one with more experience (> 3 years) became the coach. After role assignment the participants were led to two different rooms respectively in our department building, and the tasks were explained to them step by step. They were both told to complete the tasks as fast and accurately as they could. Then they were given an opportunity to get comfortable with the environment and talk with their partner at the other end via a Google Talk VoIP channel. In the experiments, they communicated through the same audio channel and shared the visual context of the student's space.

Depending on the view configuration, the coach was told whether he/she could control the view and how to control the view if possible. In multi-2D view mode, the coach was shown three radio buttons to select three views respectively. In the 3D view mode, the coach was shown how to move and drag the mouse in the application window to change the viewpoint arbitrarily in the scene. The coach was then given several minutes to practice controlling the view.

3.3 Workspace and Tasks

The video-mediated system was set up in two remote rooms in a department building.

Student's Workspace: Figure 2(a)-(c) shows the student's work space which consisted of an array of cameras and a 61-inch NEC plasma display. Based on different camera conditions (fixed view, multi-2D view, 3D view), different cameras were used to capture the student in the scene. The video streams were then aggregated and transmitted to the coach's site, and were also rendered locally on the large plasma display to show a mirrored view for the student, thereby providing the shared visual context.

Coach's Workspace: Figure 2(d) shows the coach's space with a renderer computer which received the video data from the student's site and rendered them on a 17" Dell desktop LCD display. The coach could see the student's live video on the screen, and could also change his/her viewing perspective of the cyber-space with a mouse and a keyboard.

Materials: A life-size virtual basketball hoop (Figure 2(e)) was rendered in the scene which could be seen by both the coach and the student. Two toy plastic balls with a diameter of 6 inches were used by the student. Both the coach and the student were wearing a headphone with microphone, thus could talk via the VoIP connection.

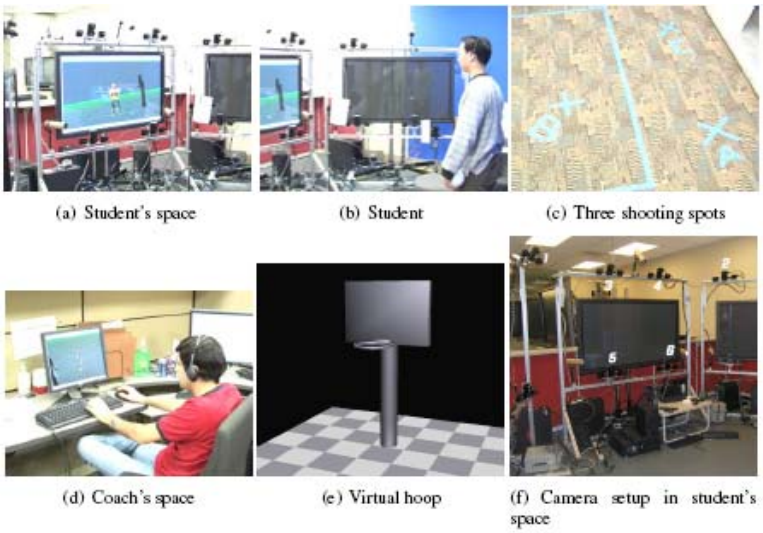


Figure 2: Workspaces

Tasks. The coach instructed the student by audio to learn basic basketball skills in two steps: (1) move the ball on top of the virtual basket, and then drop it into the basket; (2) stand at three positions (i.e., A, B, and C, respectively, as marked on the laboratory floor, Figure 2(c)), and attempt to shoot the ball into the virtual basket. When shooting the ball, the coach corrected the student's pose by shaping his/her arms, hands, shoulders, knees, etc.

3.4 Camera Conditions

As mentioned before, the learning tasks were performed under three camera conditions: fixed view, multi-2D view, 3D view, respectively. The order of the three sets is randomized to minimize the noise from learning effect.

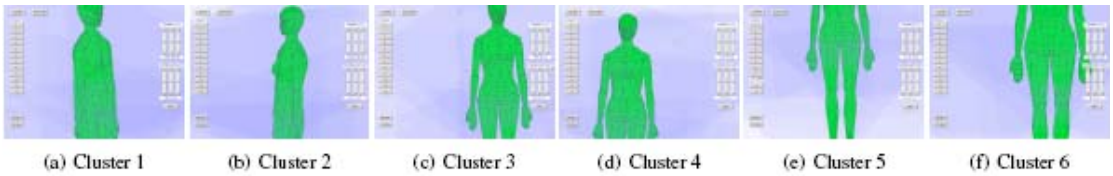


Figure 3: Simulated views of each camera cluster

Figure 2(f) shows the camera setup in the student's space. We had six camera clusters in total, Cluster 1-6 as marked in Figure 2(f). Each cluster consists of three black-and-white cameras (bottom) and one color camera (top). We use "Camera *i*" to indicate the color camera in Cluster *i*. For example, Camera 3 denotes the color camera in Cluster 3. Figure 3 shows the simulated view of each camera cluster. Below we describe the camera setup for different visual representations.

- *Fixed view*: Only one 2D camera (Camera 3) was used, so the coach could see the scene from a fixed perspective.
- *Multi-2D View*: Three 2D cameras - one on the left of the student (Camera 1), one on the center (Camera 3), and one on the right (Camera 2), but all in front of him/her- were used in this setting. The coach was thus able to switch among three corresponding viewpoints: *left*, *center*, and *right*, by clicking three radio buttons on the renderer interface, respectively. The viewpoints in this scenario were "physical", for they were generated naturally from the cameras.
- *3D View*: Six 3D camera clusters (i.e., twenty four 2D cameras) were used to reconstruct the 3D model of the scene. The coach could then drag the mouse freely in the renderer to observe the scene from any viewpoint. The viewpoints in this scenario were "virtual" because they were simulated by reconstructing a 3D model from multiple 2D video streams. The viewer's eyes can be represented by a virtual camera. Changing the perspective of the viewer is equivalent to moving the virtual camera in the 3D scene.

Since the student could hardly change the view when performing physical movement, the perspective remained unchanged as a mirror view across all configurations, which came from Camera 3.

3.5 Hypotheses

We made two sets of hypotheses, with regard to task performance and user control pattern.

Task Performance. We hypothesized that both the multi-2D view and 3D view would improve performance over fixed view, because more perspectives would increase situation awareness and conversational grounding to facilitate the mutual understanding between the participants. 3D view should further outperform multi-2D view because it provides much more perspectives, and thus richer information of the scene. Nevertheless, we expected that the task completion time in fixed view to be shorter than that of multi-2D view and that of 3D view, because fixed view did not involve any user input as there was only one view available to the coach.

Control Pattern. (a) For multi-2D view, we anticipated that the three views (left, center, right) would have different selection statistics, because they provided visual cues that differ in effectiveness. Based on prior work for user customization [Yu 2006], we also expected different user preferences on view selection in multi-2D view even for the same collaborative task. (b) As for 3D view, we hypothesized that the selected views would

not be evenly distributed over the 3D scene space. In particular, we expected that most of the time users would observe the scene from a few number of positions. Furthermore, we anticipated that the task completion time in 3D view would be longer than that of multi-2D view, because rotating the view in 3D using mouse takes more time than clicking the radio buttons to select a 2D view.

3.6 Analysis

Numerical Analysis. Any view change in each session was recorded in numerical values for analysis. In multi-2D view mode and 3D view mode, all view changes were logged along with the timestamps.

Videotape Analysis. Each session at the coach's space was recorded by a camcorder. The videotape was later screened to compute the task completion time by each pair of participants. The conversations between the coach and the student were used for qualitative analysis. Although the view switching/moving choices were all logged numerically, the visual progress recorded in the videotape facilitates our identification of control patterns by users.

Video Data Analysis. All videos of the scene across three configurations were recorded. Each stream was saved as a data file. Since 2D video was only taken at coach's space, 3D video became very helpful for analyzing the behavioral patterns of students.

4 Results

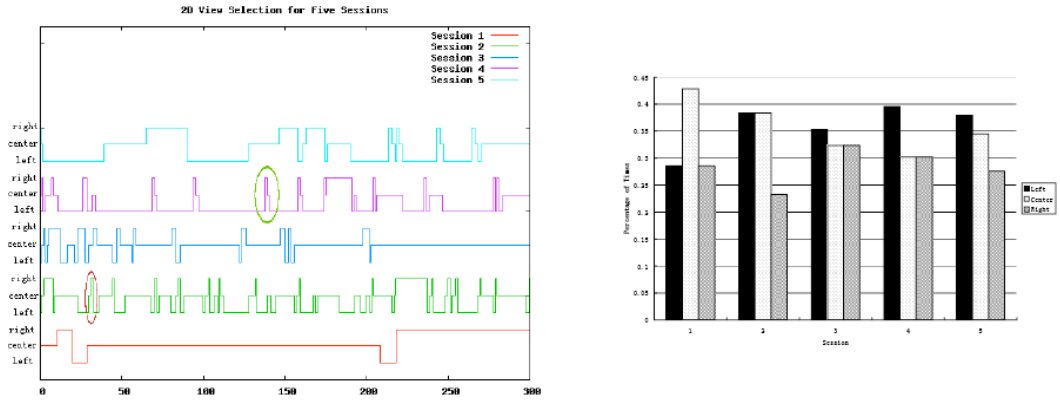
We present the experimental results in this section, along with a discussion about the design implications for future video-mediated learning systems. The first pair of participants was not able to complete the task, and was thus withdrawn from the data set.

4.1 Control Patterns

We were interested in how users controlled the view in different configurations and whether there were certain patterns when they switched the view. Results of this section were from numerical analysis of logged data for both multi-2D view and 3D view configurations.

4.1.1 2D View

Figure 4(a) shows the view switching pattern by different users in multi-2D view mode. Interestingly, we noticed how frequent users just quickly "browsed" through three views instead of selecting the favorite view with intention. The circle in Session 4 in Figure 4(a) shows a pattern that the user switches from right to center and then to left, and the circle in Session 2 shows a pattern in the reverse way. It is clear that there were a number of such patterns in each session. Since there were only three views to choose from, it was relatively easy for the user to go through all views quickly. This is also reflected in Figure 4(b) where each view was selected for similar numbers of times. This observation does not support our hypothesis that "the three views would have different selection statistics". However, it might be an interesting finding on user behavior, because it evokes questions such as how many views should be provided to users (to strike a good balance between the richness of visual information and the level of distraction), and how frequent each view should be switched on to refresh the user's memory about the particular view.



(a) View switching log (b) Percentage of time that each view was selected

Figure 4: Multi-2D View Results

We computed the percentage of time each view was selected for all sessions. It turns out that on average the left view was selected 33% of the time, the center view 44% of the time, and the right view 22% of the time in a session. Through the analysis of videotape, we found that right view was least useful because the view of the user was obstructed by the basketball backboard under this configuration. This result suggests the importance of occlusion detection and avoidance in automatic view control.

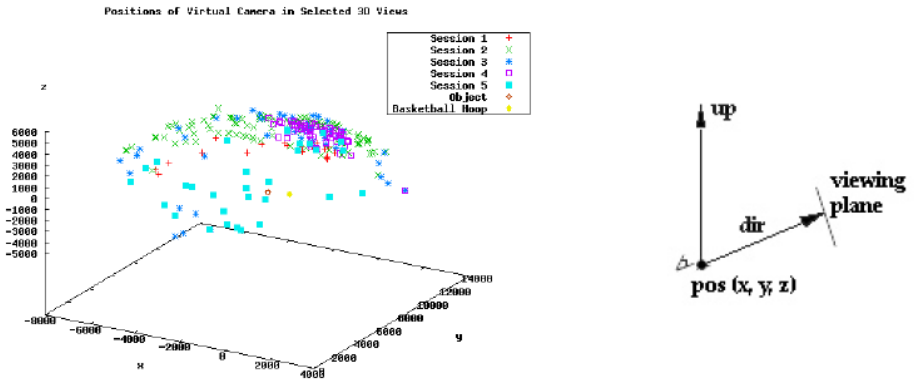
From Figure 4(a) we noticed a split in user behavior on view selection. In particular, two extreme cases appear to be most interesting. The user in Session 1 (User 1) seldom changed the view at all. In the 227-second session he only switched view for 5 times. In contrast, the user in Session 2 (User 2) changed the view very frequently. In the 299-second session she/he made 73 switches in total. More specifically, for 83% of the time User 1 stayed on center view while for only 9% of the time he chose left view and the same with right. Interview with this particular user reveals that he decided to stick to center view after a comparison of the three available views, because he thought center was almost the only view that provided him enough visual information to make a judgment. On the other hand, User 2 stayed on left, center, right view for 36%, 43%, and 21% of the time, respectively. We also noticed an outstanding number of "quick scanning" behaviors for User 2. This result provides evidence in support of our hypothesis about user preferences, and suggests that user customization is indispensable for automatic view control system. Different users might have radically different preferences on how views should be selected.

4.1.2 3D View

As mentioned above, changing the perspective in a 3D space is equivalent to moving a "virtual camera" in the space. Figure 5(b) shows the model of the virtual camera. Figure 5(a) shows the positions of the virtual camera (*pos*) in 3D view selected by different users. The *up* vectors of all virtual cameras were assumed to be upward. The *dir* vectors were not drawn because the readers can imagine they were pointing from the positions of the virtual cameras to the student in the scene. Notice that the sixth series in Figure 5(a) indicates the position of the student in the 3D space, and the seventh series shows the position of the virtual basketball hoop.

We hypothesized that the positions of selected views would not be evenly distributed over the 3D scene space. This hypothesis is confirmed by the data shown in Figure 5(a). There is clearly a "clustered" area of view positions on the upper right part of the 3D space. Those positions were roughly 45-degree above the observed objects (i.e., the student and the basketball hoop). This is partly due to the nature of the tasks in our study, because that angle was ideal for the coach to check if the ball would go into the basket. This result also provides support for our argument that the "optimal" view configuration in a video-mediated environment depends

largely on the nature of the collaborative tasks.



(a) Positions of virtual camera in 3D view (b) Virtual camera model

Figure 5: 3D View Results

In 3D view, users still show different preferences on view control, which further confirms the need of user customization. For example, the user in Session 5 was the only one who spent the majority of the time observing from the bottom of the horizon (with $z \leq 0$) in the 3D scene, and the user in Session 2 appeared to favor the left half of the upper hemisphere (with $x \leq 0$) while others did not.

We also noticed that a large portion of view movement occurred within small distances. This might indicate that smooth and small movement of virtual camera would likely be favored over abrupt and large movement of the camera.

Table 1: Learning performance in completion time and number of hits (out of seven shots)

	Completion Time (sec)		Number of Hits	
	Mean	SD	Mean	SD
Fixed View	163.75	27.58	1	1
Multi-2D View	215	36.68	2	1
3D View	225	8.04	5	1

4.2 Performance

Table 1 summarizes the learning performance across three camera conditions as measured in task completion time and number of hits made by the student out of seven shots. An ANOVA (analysis of variances; $F[2; 9] = 5; 968; p \leq 0:03$) suggests that the result is statistically significant.

4.2.1 Completion Time

First, we expected difference in task completion time across three camera conditions, that is, the visual representation provided to the users should have impact on the task completion time in remote collaboration. This hypothesis is proved true preliminarily in Table 1.

We hypothesized that the task completion time in fixed view would be shorter than that of multi-2D view and that of 3D view, because fixed view did not involve any user input and because only one view was available to the coach. The experimental data provides evidence in support of this hypothesis. Table 1 shows that the task completion time in fixed view is 23.8% shorter than that of multi-2D view and 27.2% shorter than that of 3D view.

Moreover, we anticipated that the task completion time in 3D view would be longer than that of multi-2D

view, because rotating the view in 3D using mouse takes more time than clicking the buttons to select a 2D view. It turns out that the difference is small (4% as Table 1 shows). This is partly due to the fact that users tended to "browse" through three views in the multi-2D view mode, which took more time than simply selecting the favorite viewpoint (as expected).

4.2.2 Learning Quality

During the learning, one shot was made by moving the ball right above the basket and drop it, and two shots were made on each spot, A, B, and C, respectively. Table 1 shows the numbers of hits the students achieved on average out of the seven shots. An ANOVA analysis ($F[2; 9] = 19{:}63$; $p \leq 0{:}001$) indicates that this result is statistically significant.

We hypothesized that both the multi-2D view and 3D view would improve performance over fixed view. We further hypothesized that 3D view would achieve the best performance in learning. According to Table 1, these hypotheses are proved to be true with significant probability.

We note that the task completion quality measured in number of hits under 3D view was considerably higher than the other two, by a factor of 5, and 2.5, respectively. This result demonstrates the strong potential of 3D visual information in facilitating sport learning. With 3D visual representation, the users can observe the remote scene from arbitrary angles, which gives much richer context information than the 2D representation. The coach can, for example, check the student's pose from various perspectives rather than in flat images, which proves to achieve more effective training/learning results.

In addition, we did not find substantial evidence to suggest an obvious advantage of multi-2D view over fixed view in terms of learning quality. This is partly due to our selected fixed view (center) captured the front of the student, and also played an important role in multi-2D view. The other two views in multi-2D mode, however, did not appear to be more helpful than the center view. Therefore, careful placement and selection of 2D cameras has a strong impact on the effectiveness of multiple 2D views.

4.3 Questionnaire Response

After each experiment, we gave questionnaires to the participants. We asked about the effectiveness of visual and audio links to facilitate remote interaction. In the multi-2D view mode, we asked the users to rank the three given views in the order of usefulness. In 3D view mode, we asked whether the free view change was helpful.

The effectiveness of the shared visual and audio links between pairs of participants was confirmed by the mean scores of 4.33 and 4.83 (both out of 5), respectively.

For fixed view, all coaches thought having the view fixed hindered the collaboration with the students with a mean score of 4.6 out of 5. For multi-2D view, 80% of the coaches voted "center" as the most useful view, whereas the other 20% voted for the left view. For 3D view, the students thought being able to change the view freely helps collaboration with a mean score of 4.8 out of 5.

Many participants left either verbal or written comment saying that they actually had fun in the experiments. As pointed out by Mueller *et al.* [Mueller et al. 2007], video-mediated systems are very promising in the domain of sports.

5 Conclusion

We investigated the impact of visual representation under three camera conditions: *fixed view*, *multi-2D*

view, and *3D view*. Furthermore, we identified several interesting control patterns by logging the user actions in the experiments. Our study shows that the visual context has significant impact on the performance of video-mediated learning tasks. It also demonstrates the strong potential of 3D visual information in facilitating sport learning tasks.

References

H. Baker, N. Bhatti, D. Tanguay, I. Sobel, D. Gelb, M. Goss, W. Culbertson, and T. Malzbender. (2005) Understanding performance in coliseum: an immersive videoconferencing system. *ACM Transaction on Multimedia Computing, Communications, and Applications* (pp. 190-210).

M. Chen (2001). Design of a virtual auditorium. *Proceedings of the 9th annual ACM international conference on Multimedia* (pp. 19-28).

S. Fussell, R. Kraut, and J. Siegel (2000). Coordination of communication: Effects of shared visual context on collaborative work. *Proceedings of Conference on Computer Supported Cooperative Work* (pp. 21-30).

S. Fussell, L. Setlock, and R. Kraut (2003). Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. *Proceedings of Computer-Human Interaction* (pp. 513–520), New York: ACM Press.

W. Gaver, A. Sellen, C. Heath, and P. Luff (1993). One is not enough: Multiple views in a media space. Proceedings of the conference on Human factors in computing systems (pp. 335-441), New York, ACM Press.

S. Jung and R. Bajcsy (2006). Learning physical activities in immersive virtual environments. *Proceedings of the 4th IEEE International Conference on Computer Vision Systems*, New York City, January 5-7th, 2006.

R. Kraut, S. Fussell, and J. Siegel (2003). Visual information as a conversational resource in collaborative physical tasks. *Human-Computer Interaction* 18 (pp. 13–49).

A. Mankin, L. Gharai, R. Riley, M. Maher, and J. Flidr (2000). The design of a digital amphitheater. *Proceedings of Network and Operating System Support for Digital Audio and Video (NOSSDAV).*

Microsoft Conference XP: http://research.microsoft.com/conferencexp/.

F. F. Mueller and S. Agamanolis (2007). Sports over a distance. *Personal and Ubiquitous Computing* 11 (pp. 633-645).

A. Ranjan, J. P. Birnholtz, and R. Balakrishnan (2006). An exploratory analysis of partner action and camera control in a video-mediated collaborative task. *Proceedings of the 20th anniversary conference on Computer supported cooperative work* (pp. 403–412), New York, ACM Press.

H. Towles, W.-C. Chen, R. Yang, S.-U. Kum, H. Fuchs, N. Kelshikar, J. Mulligan, K. Daniilidis, L. Holden, B. Zeleznik, A. Sadagic, and J. Lanier (2002). 3d tele-collaboration over internet2. *Proceedings of the International Workshop on Immersive Telepresence*, Juan-les-Pins, France.

E. S. Veinott, J. Olson, G. M. Olson, and X. Fu (1999). Video helps remote work: speakers who need to negotiate common ground benefit from seeing each other. *Proceedings of Computer Human Interaction* (pp. 302-309).

Z. Yang, B. Yu, W. Wu, R. Diankov, and R. Bajcsy (2006). A study of collaborative dancing in tele-immersive environment. *Proceedings of the 8th IEEE International Symposium on Multimedia (ISM'06)*, San Diego, CA.

B. Yu (2006). MyView: Customizable Automatic Visual Space Management for Multi-STream Environment. *Ph.D. thesis*, University of Illinois at Urbana-Champaign.

Acknowledgements

This research is supported by National Science Foundation (NSF) under grants CNS-0448246, CMS-0427089, NSF SCI 05-49242 and NSF CNS 05-20182. The presented views are those of authors and do not represent the position of NSF.