Presiding Over Accidents: System Direction of Human Action

Bibliographic Reference:

Jeffrey Heer, Nathaniel S. Good, Ana Ramirez, Marc Davis, and Jennifer Mankoff. “Presiding Over Accidents: System Direction of Human Action.” In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2004) in Vienna, Austria*. ACM Press, 463–470, 2004.

Presiding Over Accidents: System Direction of Human Action

Jeffrey Heer[1], **Nathaniel S. Good**[2], **Ana Ramirez**[1,2], **Marc Davis**[2], **and Jennifer Mankoff**[1]

[1]Group for User Interface Research
Computer Science Division
University of California, Berkeley
{jheer, anar, jmankoff}@cs.berkeley.edu

[2]Garage Cinema Research
School of Information Management and Systems
University of California, Berkeley
{ngood, marc}@sims.berkeley.edu

"[...] the terrible burden of the director is [...] to select from what happens during the day which movement shall be a disaster and which a gala night. His job is to preside over accidents." - Orson Welles

ABSTRACT

As human-computer interaction becomes more closely modeled on human-human interaction, new techniques and strategies for human-computer interaction are required. In response to the inevitable shortcomings of recognition technologies, researchers have studied mediation: interaction techniques by which users can resolve system ambiguity and error. In this paper we approach the human-computer dialogue from the other side, examining system-initiated direction and mediation of human action. We conducted contextual interviews with a variety of experts in fields involving human-human direction, including a film director, photographer, golf instructor, and 911 operator. Informed by these interviews and a review of prior work, we present strategies for directing physical human action and an associated design space for systems that perform such direction. We illustrate these concepts with excerpts from our interviews and with our implemented system for automated media capture or "Active Capture," in which an unaided computer system uses techniques identified in our design space to act as a photographer, film director, and cinematographer.

Author Keywords

direction, recognition, error, mediation, active capture, error-prone systems, multimedia systems design

ACM Classification Keywords

H.1.2 [**Models and Principles**]: User / Machine Systems; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces.

INTRODUCTION

Despite its natural feeling, everyday human-to-human interaction is far from an error-free process. People both consciously and unconsciously engage in techniques to ensure mutual understanding. For example, people commonly use body language or expressions such as "huh?" or "what?" to indicate confusion or missed information, and backchannel utterances like "ok" and "uh huh" to signify comprehension [7].

Supporting "natural" human-computer interaction imposes additional challenges given the limited capabilities of input technologies such as speech recognition and computer vision. As a result, researchers have developed a body of work on *mediation*: interaction techniques for the resolution of system ambiguity and error. Mediation techniques have a rich history in the speech recognition literature [16], and have recently received systematic exploration and toolkit support in the GUI domain [17].

Though such techniques have proven valuable for bridging the gap between human and machine understanding, the converse of these techniques—direction and mediation of human performance by the computer—has received less extensive treatment within HCI circles. System-initiated direction of human action makes possible a rich and compelling domain of applications. Examples include automated media capture [8,9,10], automobile anti-sleep systems [2], emergency evacuation systems for buildings, and even automated trainers that instruct Tai-Chi [5] or improve a user's golf swing. It is important that researchers and developers have theoretically and practically grounded strategies for creating such systems.

To better design computational systems that guide human performance, we first wanted to understand how *humans* guide human performance. We sought domains in which people were instructed on the performance of physical actions. We interviewed film and theater directors, as well as a children's portrait photographer, to study cases where the instructor is trying to generate a specific emotional and physical response from the actors or participants. We interviewed golf and aikido instructors for cases where participants are learning new patterns of movement. We also looked at people who had to give instructions over the telephone, such as 911 operators and triage nurses, as their

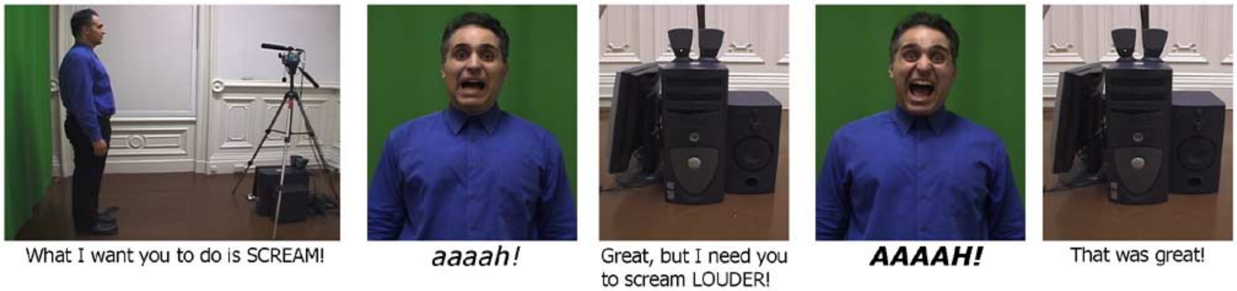

What I want you to do is SCREAM!
aaaah!
Great, but I need you to scream LOUDER!
AAAAH!
That was great!

Figure 1: Example of Active Capture process for a Scream

work contexts share with computer systems the need to direct and monitor action over a limited sensor channel.

Based on these interviews and a review of prior work, we propose a design space and direction strategies for systems that guide physical human action, taking the first steps towards translating the skills of human practitioners into applicable strategies for human-computer interaction. Throughout the paper we use automated media capture, or "Active Capture," as a running example of an implemented human-directing system.

ACTIVE CAPTURE

Active Capture is a new paradigm for media capture in which, by applying media production knowledge, multimedia output, and computer vision and audition, the unaided computer can act as a photographer, film director, and cinematographer [8,9,10]. From the system's perspective, the goal of Active Capture is to capture reusable, annotated media content in a completely automated fashion. From the user's perspective, the goal is to provide enjoyable new experiences through which everyday users can become producers (and stars!) of media.

Figure 1 shows our implemented Active Capture "Scream" scenario, in which the computer uses audio prompts to direct the user to scream into the camera. The system compares the captured scream's volume and duration against pre-set threshold values to determine the scream's acceptability. The captured scream shot can then be automatically incorporated into adaptive media templates [8] to generate any number of short personalized videos. Implemented templates include scenes and trailers from the films Godzilla, Blair Witch Project, and Terminator 2, and commercials for 7Up and MCI (Figure 2).

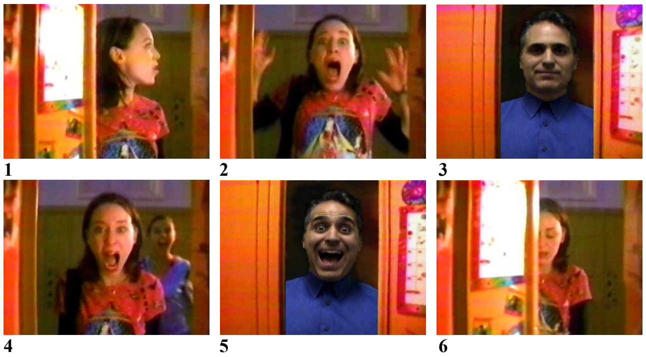


Figure 2: Stills from automatic 7Up commercial using the captured scream shot.

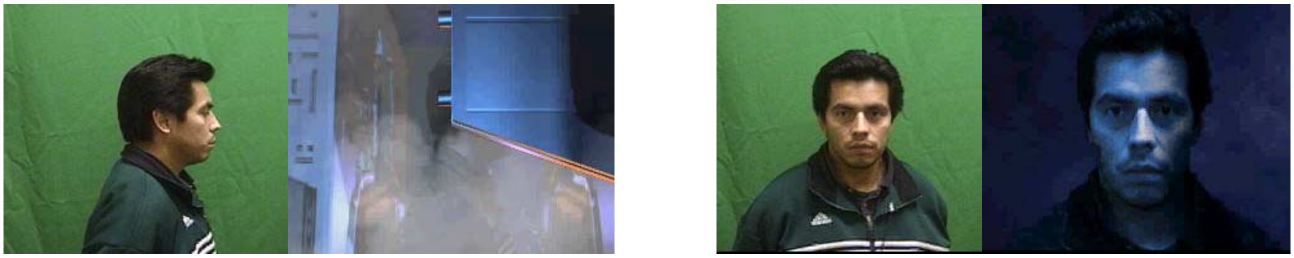
We have also implemented the Active Capture "Head Turn" scenario, in which the system uses audio/visual cues to direct the user to perform a head turn. The user is instructed to stand on marks on the floor, and look away from the camera. After motion detection is used to ensure the user isn't moving, the user is asked to slowly turn to face the camera. Eye detection routines are used to ensure the user is indeed facing the camera at the end of the turn. The resulting shot can then be used to automatically construct a personalized trailer for the film Terminator 2, in which the user appears as a killer cyborg from the future (Figure 3).

While these scenarios can be seen as an extension of automated photo booths into the realm of automatic personalized motion pictures, Active Capture, and its associated interaction design, encompasses a wider range of applications in which humans and devices work together to capture high-quality annotated media assets (*e.g.,* media messaging and greetings, photo ID, travel documentation, entertainment, marketing, and advertising).

Obviously, there is much that can go wrong in these interactions. Not only might the system's recognizers make incorrect inferences, the participant may misunderstand the system's direction or give performances that do not meet the programmed requirements for the shot. As a result, the system must adopt strategies for directing the user and provide appropriate feedback to shape the desired performance. As we plotted out strategies by which to improve Active Capture, we realized that a more thorough investigation would benefit not only the design of Active Capture scenarios, but that of any application in which a computer system could be used to automatically capture, analyze, and provide corrective feedback to physical human action. The lack of readily available guidelines for the design of such systems fueled our interest in human direction and mediation techniques.

RELATED WORK

Our work on direction techniques draws from prior work on mediation, much of which has been conducted in the fields of speech recognition and multimodal interfaces. Ainsworth and Pratt [1] and Baber and Hone [3] identify the types of



Beginning of head turn

End of head turn

Figure 3: Pictures of an Active Capture participant performing a head turn. The figure shows both the original captured footage and corresponding images from an automatically generated Terminator 2 trailer.

speech recognition errors and introduce and evaluate mediation strategies for resolving them, laying out the primary mediation strategies of *repetition* (the user repeats their input) and *choice* (the user selects from a list of possible interpretations). Yankelovich *et al*. [20] present an advanced speech system with error-correction support, in which they introduce the strategy of *progressive assistance*, providing increasingly informative assistance messages in the face of repeated error. The speech UI firm TellMe also employs a rich set of mediation strategies, including *freshness* (avoiding repeated utterances) and *graceful failure* (offering natural exits for the user, *e.g.*, time outs). Mediation work has also been done in the multimodal domain, where *modality shifts* are used to disambiguate recognition [18,19]. For example, Suhm [19] convincingly demonstrates the use of handwriting to quickly correct speech recognition errors. Mankoff *et al*. review past work on the mediation of recognizer errors [16] and provide support for mediation in a GUI toolkit [17] that focuses on interaction techniques supporting choice mediation.

Numerous projects have also examined how to structure the dialogue between computer systems and users. Brennan and Hulteen [6] describe an approach to conversation management rooted in Clark and Brennan's work on the psychology of conversation [7] and provide examples of *positive feedback* and *negative feedback* as natural techniques for detecting the need for mediation and establishing *grounding* between conversation participants. Video games also have a history of incorporating dialogue. Jellyvision, Inc., makers of the popular video game "You Don't Know Jack," have published a set of guidelines [12] for establishing an immersive (albeit one-way) conversational experience.

A few researchers and practitioners have focused specifically on molding human action and behavior. One early pioneer is Zoltan-Ford [21], who investigated techniques to facilitate human adaptation to natural language systems, controlling vocabulary and *discourse level* to shape a user's vocabulary to match that of the system. Another important example is intelligent tutoring systems [13], which use cognitive models of learning to provide guidance and feedback to students.

Direction of human performance by other humans, however, has been studied and practiced for centuries. In addition to our interviews, we reviewed manuscripts on practice, including texts by well known film directors [14, 15]. Also relevant is a rich body of psychological and educational literature on learning, including the study of different learning styles and strategies [11].

CONTEXTUAL INTERVIEWS

To better design computational systems that guide human performance of physical action, we wanted to understand how *humans* perform such direction. In deciding who to interview, we sought to cover a broad range of activities applicable to automated systems. The directors and photographer we interviewed elicit specific physical and emotional responses from subjects. Physical trainers, such as the golf and aikido instructors we interviewed, instruct and evaluate new patterns of movement that are often quite complex. Finally, the 911 operator and telephone nurse we interviewed have to monitor, persuade, and direct subjects within a limited communication medium.

Guided by contextual inquiry practices [4], we designed interviews consisting of three phases: a standard interview regarding domain knowledge; where appropriate, an observer-participation phase in which one interviewer directly participates in direction from the expert; and finally, a debriefing phase in which we elicited feedback on our own analysis, confirming that our analysis had captured the essence of their direction style. Findings from our interviews are presented below; we systematize these findings later in our Direction Strategies section.

Film and Theater Directors

We learned that film and theater directors try to *engage* their actors in a scene by guiding or pushing them to express a desired emotional state. The film director described a movie as an "emotional symphony" where each actor's part needed to be orchestrated correctly to create a coherent vision and story. Creating the right kinds of *internal* and *external motivation* was very important. The film director would try to mold actors' behavior by modifying his own behavior for a given scene and sometimes the whole day. If the film director wanted an actress to be angry, he might act angry around her all day!

The theater director tried to get actors to associate the scene with parts of their own life to make it more realistic. In one example, he wanted an actress to be "horrified," but after repeated attempts, "it wasn't exactly what [he] wanted" so he attempted to *make* her experience and express a desired emotion. While she was talking, he would "get in her face" and "say really disgusting things" to "get her to physically react how [he] wanted."

Children's Portrait Photographer

The children's portrait photographer taught us that by using *external aids* he could *make* children perform the desired action. The children's portrait photographer was interested in getting toddlers and infants to smile or pose a certain way in order to get the best picture possible. Typically, this required the help of at least two people for infants and children under 3 years-old. The photographer and the assistant would use props such as stuffed toys, bells, whistles, and balloons to get the children to look a certain direction. Getting infants or children to smile was another challenge that required multiple techniques. Making funny faces, playing peek-a-boo, or making funny noises were used for infants and smaller children. For older children, imagination and *internal motivation* played a key part in getting them to act a certain way. For example, if the photographer wanted a big smile, she could say "Imagine you have a big plate of chocolate all to yourself" to get the child to light up and smile.

Golf Instructor

Whereas directors and photographers are interested in getting a particular scene or shot, instructors such as the golf and aikido instructors that we interviewed are more interested in instilling new habits and movements over a longer period of time. To do so, the golf instructor would *decompose* complex motions into less complex parts, as well as *alternate instruction methods*, *explain consequences* of good or bad actions, and use *external aids*. The interviewed golf instructor described breaking down the complexity of a golf swing into individual parts to "sequentially build the foundation" of good practice. He decomposed the golf swing into 5 separate parts, and had students work on each part in order. He utilized a technique of *telling* the student what was correct form, *showing* what was correct, and then allowing the student to try out the technique while he observed and commented.

When students had errors in their technique, he would also frequently demonstrate what they were doing wrong so that they could see it for themselves, describe how the mistakes were affecting the outcome, and then *make* the student perform the correct technique. For example, if a student was holding a club with the face open, he would show the student how that was happening, explain how this ended up in a slice, and physically correct the student's grip and stance so that they could experience for themselves what a correct position felt like. He also used external aids, such as clubs and golf balls with lines on them to improve aim, and clubs with hinges that bent when they were swung incorrectly, to help students get continuous feedback.

Aikido Instructor

Similar to the golf instructor, the aikido instructor both described desired actions and decomposed them into pieces that could be taught and practiced individually. He described a "diagnostic space" of problems that he would notice and individually address—an important aspect was determining which aspects of technique students could change consciously and which they couldn't. Repetition of decomposed actions was key to unlearning "bad" physical habits. Unlike the golf instructor, there was a stronger focus on *internal motivation*, for example "imagine your arms filled with a large ball of ki [energy]" before attempting a forward roll. The aikido instructor also made an effort to engage *multiple methods* and *modalities*: describing a technique and its underlying philosophy, demonstrating the technique, and actually applying the technique.

911 Emergency Operator

911 emergency operators have developed means of negotiating emergency instruction over the phone when it is not quite clear what is happening at the other end of the line. They demonstrated *anticipation* of common errors and *decomposition* of complex actions into simple steps. They use both a script that describes the course of action for various problems (CPR, child birth, infant choking), as well as their own experience and intuition about common problems (*e.g.*, having a pillow under someone's head while trying to perform CPR).

In addition, they demonstrated the importance of *confirming* each step in the process, as well as the method of *backtracking* for finding and fixing possible errors. A typical scenario for 911 operators is to instruct the caller in CPR. Because the operator has no way of seeing if this is being done properly, and is often instructing CPR novices, they must decompose the act into explicit sequential instructions, often reconfirming the actions and outcomes at each step. Confirmation is frequently achieved by using the word "OK?", for example, "I want you to get the phone as close as possible to him, OK?" Sometimes they phrase commands in the form of questions such as "Tilt his head towards the ceiling?" to simultaneously tell what action to perform as well as ask if the action has been performed correctly. Backtracking is useful for correcting errors that come up in the call. In backtracking, the operator returns to a previous state that she is certain was carried out successfully, and attempts to work forward from there. This is especially important in CPR, where the proper execution is dependent on successful execution of the previous steps.

911 operators also exhibited varying the *level of discourse* and tone of language (or operator's *impression*) to instruct strangers over the phone in complex life or death situations. The 911 operator we interviewed found it was important to take charge of a call after getting the necessary information

from someone on the phone. The switch from inquisitor to instructor required a shift in language and tone. Phrases changed from "What is your X?" to "I want you to do X."

Telephone Triage Nurse

Unlike the 911 operators who briefly speak to callers in one-time emergency situations, the telephone triage nurse we interviewed frequently deals with repeat callers, (over 50%), her calls are longer (10 minutes on average) and she provides instructions on less critical but important matters, such as how to take a baby's temperature. Because these calls are more informative and less likely to be emergencies, the tone of voice and style of conversation is more accommodating and less forceful than 911 operators, and humor and empathy are often used. Discussing a recent experience teaching a mother how to take a baby's temperature, the triage nurse described her role and methods as "I wasn't telling her, I was more asking her kinds of questions, so I wanted her to feel a part of the process. Not to feel that 'I'm going to tell you how to do this,' but that she was accomplishing something at the same time." It was important to build a foundation of trust and accomplishment by *engaging* the caller and not criticizing them. The nurse also described dynamically changing her *level of discourse* (sometimes "dumbing it down") until she felt the caller understood her.

Summary of Direction Techniques from the Interviews

Our interviews uncovered numerous strategies employed by experts to guide specific human actions. The interviews covered a diverse set of instruction techniques and contexts, which helped us discern similarities and differences among them. In the case of the directors and the photographer, we learned that external aids as well as internal and external motivation can help us direct people to perform a given action. While teaching people new actions, the golf and aikido instructors applied many of these same techniques but also decomposed complex actions into smaller more easily learnable parts. When faced with a limited communication channel, the lessons of the 911 operators and the triage nurse are also illuminating. From 911 operators, we learned that taking charge of the call, assertive language, confirmation, and backtracking for error correction are important when asking people to perform specific actions in a short time frame. From the triage nurse, we saw the addition of more varied levels of discourse, empathy, humor, and establishment of rapport with the caller in the face of reduced time pressures.

DIRECTION STRATEGIES

Analyzing the results of our interviews revealed a rich set of direction strategies used in diverse contexts. Although the practices of our interviewees were often highly nuanced, involving various social and problem-solving skills, repeated patterns did emerge. These strategies are motivated by the need for *grounding*—establishing mutual understanding between participants [7]—and progress towards a specific set of *goals*—guiding the subject to a desired outcome. In this section, we describe successful direction strategies for achieving grounding and goal-directed progress observed in our interviews and review of prior work. We segment these strategies into three classes: general *design strategies* pertinent to an entire interactive session, individual *direction and feedback strategies* for guiding the subject, and *mediation strategies* for resolving complications and errors.

Design Strategies

Anticipation: Anticipate common errors before they happen. Actively seek out these problems and address them before they disrupt the interaction.

Appropriate System Impression: Adopt the appropriate tone and role for the context of the interaction. Urgent tasks may necessitate a curt and insistent style, while recreational contexts afford a more relaxed and flexible tone.

Direction and Feedback Strategies

Decomposition: Break down complex actions into a series of simpler sub-actions. Decompose troublesome actions into sub-parts, identify those parts that are causing difficulties, and address those explicitly before re-attempting the larger, composite action.

Imaginative Engagement: Immerse the subject in the experience by engaging their emotions or imagination. Internal motivation is one avenue for accomplishing this.

External Aids: Use physical props or other external aids to guide human actions and provide implicit feedback. Examples include lines on a golf putter to assist aiming and footmarks on a floor indicating where to stand.

Confirmation: Explicitly query the subject to ensure they are in the expected state. For complex tasks it may be desirable to have the subject verbalize their understanding of the task. Even if a system cannot parse these utterances, the practice can still aid the subject's learning and memory.

Consequences: Explain the consequences, both positive and negative, of particular actions. This provides more concrete incentives to maintain beneficial actions and correct detrimental behavior.

Mediation Strategies

Freshness: Avoid repeating utterances, even when giving an instruction nearly identical to a previous one. Maintain consistent vocabulary, but don't repeat items verbatim. At a bare minimum, designers should make multiple versions of instructions that may be repeated [12].

Progressive Assistance: Address repeated problems with increasingly targeted feedback. Provide "successively more informative error messages which consider the probable context of the misunderstanding" [20].

Method Shifts: When one form of instruction fails, try another. Direction can vary between *telling* the subject what to do, *showing* them how to do it, or *making* them do it.

Modality Shifts: When a particular direction approach repeatedly fails, switch or augment the modalities of communication, *e.g.*, use visual rather than auditory cues. Changing or using multiple modalities may prove more effective to a larger audience due to the different aptitudes of auditory, visual, and kinesthetic learners [11].

Level of Discourse: Simplify the vocabulary and language structure when people are having difficulty understanding. Conversely, be concise once grounding is established.

Backtracking: When grounding is lost, backtrack to the last state of mutual understanding. By returning to a state of common ground, the system and user can then again proceed towards the goal.

Graceful Failure: When all else fails, provide the subject natural exits from the interaction. Recognize over-repetition and respond by pursuing an alternate goal or allowing the interaction to proceed to completion, despite the error.

DESIGN SPACE ANALYSIS

To better apply these various direction strategies, we found it useful to further structure the myriad design options available to human-directing systems, making clear and explicit the various routes system designs can take. The result is a design space characterized by four components—*direction*, *capture*, *analysis*, and *feedback*—each chained together in an interactive, goal-directed loop. An iteration of this loop represents a single cycle of direction, capture, processing, and feedback. It is at this level that our direction and feedback strategies can be applied. It is also important that the system keep a *memory of the interaction*. This can be used to establish grounding and enact mediation strategies by shifting options in the design space in a principled manner in response to errors. We now describe these processes in more detail.

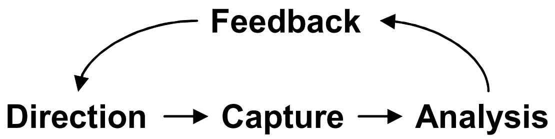


Figure 4: Design space components, chained together in a goal-directed, interactive loop.

Direction

When a system expects or requires a particular action from the user, direction is used to elicit the desired course of action. When directing a human user, the system can either *tell* the user what to do (instruction), *show* the user how to do it (demonstration), or *make* the user do it (through physical stimulus or manipulation). Note that these approaches are not mutually exclusive. For example, the Scream scenario cue "What I want you to do is SCREAM!!!" (Figure 1) uses elements of all three. Our interviewees all utilized *tell*, *show*, and *make* methods for direction, often in combination. Varying this type of direction is one way to apply both *method shifts* and *freshness* strategies.

Direction is rarely an exact specification of the desired action and in practice can be quite subtle. As discussed by the theater and film directors and the children's portrait photographer, directive acts lie on a spectrum between *internal* and *external motivation*. For example, consider the difference between the directions "Act as if your whole body is on fire" and "Flail your limbs about wildly." Internal motivation, in particular, can be used to more fully *engage* the user in the interaction. Furthermore, some directed actions may be involuntary—the response to some form of stimulus presentation—and thus do not require any forethought. As illustrated in our interviews with the children's portrait photographer and the theater director, these stimulus-response reactions can be of particular use in eliciting performances from non-actors. It is also common to use *external aids* to engage users or disambiguate direction (*e.g.*, stand on the footmarks on the floor).

After receiving direction, the user may not act right away. It may be necessary to use *triggers*, or prompts that signal the human to act. A stereotypical trigger in film direction is the shout of "Action!" Some instructions (such as the "SCREAM!!!" cue above) also serve as their own triggers.

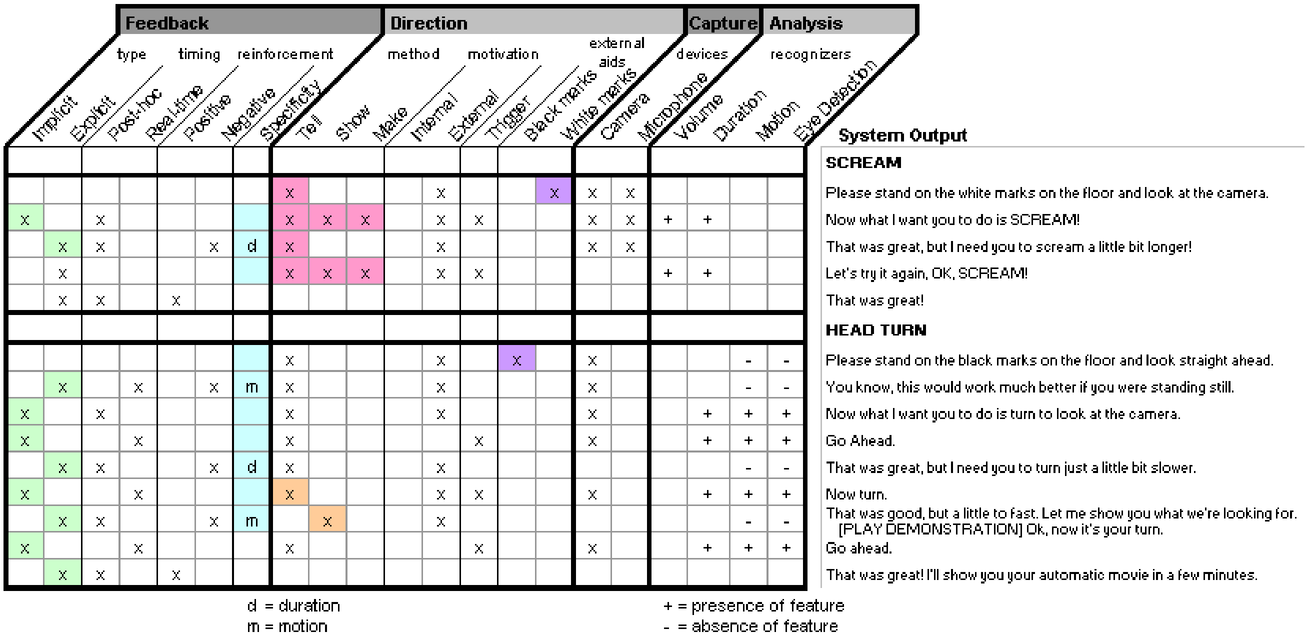
Capture

Capture concerns the mechanisms by which the system monitors and records the actions of the user and the surrounding environment. Issues to consider include the capture devices employed (*e.g.*, cameras, microphones, sensors, mouse and keyboard), their temporal and spatial configuration (*e.g.*, when and where cameras and microphones are used), and the resolution of captured data.

Though not strictly necessary in all applications, it is often important to store captured data for non-real-time analysis and as a record of the human performance. Clearly this is central to automated media capture, but can be valuable in almost any application as a means for providing feedback. For example, the golf instructor cited video as a useful tool for showing people both the highlights and problem points of their swings. This can be a useful shift in both the *method* and *modality* of feedback and serve as a means to illustrate the *consequences* of actions.

Analysis

Analysis concerns the mechanisms by which the system perceives and evaluates the actions performed by the user. Using data provided by the capture component, the system must try to determine what is being done, and how that relates to the currently active set of goals. Certainly, this requires deploying relevant technologies, such as computer vision and audition, speech recognition, or other appropriate sensor systems. More specifically, though, analysis concerns exactly how these technologies are leveraged to interpret human actions and environmental cues with respect to grounding and goal progression.



Feedback							Direction								Capture		Analysis				
type		timing		reinforcement			method			motivation			external aids		devices		recognizers				
Implicit	Explicit	Post-hoc	Real-time	Positive	Negative	Specificity	Tell	Show	Make	Internal	External	Trigger	Black marks	White marks	Camera	Microphone	Volume	Duration	Motion	Eye Detection	System Output
																					SCREAM
							x			x				x	x	x					Please stand on the white marks on the floor and look at the camera.
x		x					x	x	x	x	x				x	x	+	+			Now what I want you to do is SCREAM!
	x	x			x	d	x			x					x	x					That was great, but I need you to scream a little bit longer!
	x						x	x	x	x	x						+	+			Let's try it again, OK, SCREAM!
	x	x		x																	That was great!
																					HEAD TURN
							x			x			x		x				-	-	Please stand on the black marks on the floor and look straight ahead.
	x		x		x	m	x			x					x				-	-	You know, this would work much better if you were standing still.
x		x					x			x					x			+	+	+	Now what I want you to do is turn to look at the camera.
x			x				x				x				x			+	+	+	Go Ahead.
	x	x			x	d	x			x									-	-	That was great, but I need you to turn just a little bit slower.
x			x				x			x	x				x			+	+	+	Now turn.
	x	x			x	m		x		x									-	-	That was good, but a little to fast. Let me show you what we're looking for. [PLAY DEMONSTRATION] Ok, now it's your turn.
x			x				x				x				x			+	+	+	Go ahead.
	x	x		x																	That was great! I'll show you your automatic movie in a few minutes.

d = duration
m = motion

+ = presence of feature
- = absence of feature

Figure 5: Design space over traces of Active Capture interactions. Time proceeds downward along the table. Feedback for an action is placed directly before the next act of direction by the system. Highlighted cells show shifts in the design space corresponding to mediation strategies, such as *progressive assistance* (shifts in the specificity of feedback) and *method shifts* (between *tell*, *show*, and *make*). Notice also the interplay between implicit and explicit feedback, providing proper *levels of discourse*, and the use of disambiguating *external aids*.

For example, our Active Capture scenarios variously use volume measurement, motion detection, eye detection, and timing to evaluate a user's performance of a scream or head turn. More complex analyses are also common. Brand *et al.* use statistical learning techniques to robustly classify Tai-Chi moves [5]. Cognitive tutors [13] use production rules within a cognitive modeling environment to model student concept learning. The crucial, and often quite difficult, requirement is to decompose complex goals into sub-tasks amenable to recognition. The *anticipation* and *decomposition* strategies also necessitate analysis routines for common problems and decomposed actions.

Notice that it may be possible for users to "fool" the system—for example, our Scream scenario may recognize a loud laugh as an acceptable scream. Given the limited abilities of recognizers, such problems are in some respects inevitable. Judicious interaction design can simplify and help to interactively disambiguate the context of capture to significantly improve analysis. For example, by asking for and expecting a specific audio performance (*e.g.*, a scream) we can make use of simple and reliable audio processing. This "human-in-the-loop" approach to algorithm design effectively combines HCI and signal processing techniques to produce more robust recognizers [9].

Feedback

Feedback is the communication by the system, to the user, of the system's appraisal of the human performance and/or suggestions for improvement or modification. It is used to keep the user apprised of system state as well as to suggest ways in which the human can refine their own actions. It is tightly tied to direction—feedback and direction are often provided in the same utterance—but as a guiding concept is important to consider in its own right.

Feedback can be either *implicit* or *explicit*. Examples of implicit feedback within Active Capture are to simply move on to the next scenario, or to ask the user to repeat the previous action. The user can infer from this whether or not the previous take was acceptable, taking advantage of a higher *level of discourse*. Explicit feedback is more direct: "That was great!" or "That was great, but I need you to scream LOUDER!"

Feedback can also employ *positive* or *negative reinforcement*. Positive reinforcement rewards participants for a job well done or praises specific goals met. Negative feedback can let the user know what was done incorrectly, but need not be demeaning. For example, it was common for our interviewees to use the *consequences* strategy to explain the negative results of an action left uncorrected.

Furthermore, feedback can be supplied *post-hoc* or in *real-time*. In many cases it is sufficient to let the user know how their actions were received after the fact. Real-time feedback, however, may clarify and expedite an action while fostering *engagement*. For example, the Scream scenario could be rebuilt using a volume meter positioned next to the camera along with a target volume level.

Finally, the *specificity* of the feedback, or how targeted it is towards resolving a particular problem, is key to providing *progressive assistance*. There is a world of difference between telling a user "That wasn't good enough" and "You need to scream longer."

Mediation

In addition to these individual direction and feedback options, it is important to consider how the behavior of the design space varies over time in order to mediate errors and ambiguity. The strategies of *method* and *modality shifts*, for instance, can be understood as shifts from one set of options to another within the design space. Examples include shifts between *tell*, *show*, and *make*; between *internal* and *external motivation*; and between audio and video cues. Figure 5 presents the design space across two traces of Active Capture scenarios. The highlights in the figure illustrate the mediation strategies achieved through shifting design space options.

To establish grounding it is also crucial that the system keep a *memory of interaction*. At minimum, this memory should consist of (a) what cues the system has employed, (b) the kind and frequency of encountered errors, and (c) the user's history of successful actions and *confirmations*. Condition (a) enables *freshness*, condition (b) allows for *graceful failure*, and condition (c) is crucial to performing *backtracking*. Together, conditions (b) and (c) enable the system to suitably change the *level of discourse*.

CONCLUSION

An important class of future computing applications will be capable of sensing and directing human action. In this paper, we presented findings from contextual interviews with domain experts in various fields that seek to direct, shape, and guide physical human action. We summarized these findings in a collection of direction and mediation strategies and a preliminary design space for human-directing systems, illustrated by our implemented "Active Capture" system. The system was recently demonstrated at a major conference [10], successfully capturing 14/16 (87.5%) participants, of whom 10/14 (71.4%) required the use of mediation strategies across multiple takes.

In future work we plan to perform a deeper exploration of the identified strategies and design space. We are currently undertaking the design and implementation of additional Active Capture applications, including a picture booth for automating the creation of student ID cards, as well as devising studies to better understand when and how to apply our identified direction strategies. In practice we have observed clustering of related strategies (*e.g.*, method and modality shifts, confirmation and backtracking), but more careful study is needed to understand strategy applicability, composability, and interdependence. In so doing, it is our hope that the framework laid out in this paper will prove useful to systems and interaction designers, and that the techniques presented will lead to applications that more fully exploit the potential of human-computer interaction by supporting the direction and mediation of human action.

ACKNOWLEDGMENTS

We are indebted to all our interviewees for sharing their experiences and expertise. We also thank the members of Garage Cinema Research and the Group for User Interface Research for their feedback and insight. This work was supported by NSF grants IIS-0205644 and IIS-0209213. The first author was supported by an NDSEG fellowship.

REFERENCES

1. Ainsworth, W. A. and Pratt, S.R. "Feedback strategies for error correction in speech recognition systems", International Journal of Man-Machine Studies, 36(6), 833-842, 1992.
2. Ayoob, E., Grace, R., and Steinfeld, A. "A User-Centered Drousy Driver Detection and Warning System", Proc. of DUX 2003.
3. Baber, C. and Hone, K.S. "Modelling Error Recovery and Repair in Automatic Speech Recognition", International Journal of Man-Machine Studies, 39(3), 495-515, 1993.
4. Beyer, H. and Holtzblatt, K. *Contextual Design: A Customer-Centered Approach to Systems Designs*. Morgan Kaufmann, 1997.
5. Brand, M., Oliver, N., and Pentland, A. "Coupled hidden Markov models for complex action recognition". Proc. of CVPR 1997.
6. Brennan, S.E. & Hulteen, E.A. "Interaction and Feedback in a Spoken Language System: A Theoretical Framework", Knowledge-Based Systems 8(2-3), 143-151, 1995.
7. Clark, H.H. and Brennan, S.E. "Grounding in Communication". *Perspectives on Socially Shared Cognition*. APA Books, 1991.
8. Davis, M. "Editing Out Video Editing", IEEE MultiMedia, 10 (2). 54-64, April-June 2003.
9. Davis, M. "Active Capture: Integrating Human-Computer Interaction and Computer Vision/Audition to Automate Media Capture", in ICME 2003, Baltimore, MD, Vol. II, 185-188, 2003.
10. Davis, M. "Active Capture: Automatic Direction for Automatic Movies (Demonstration Description)." Proceedings of ACM Multimedia 2003. Berkeley, California, 88-89, 2003.
11. Dunn, R. and Dunn, K. *Teaching Students Through Their Individual Learning Styles*. Reston Publishing. 1978.
12. Gottlieb, H. "The Jack Principles of the Interactive Conversation Interface", Jellyvision, Inc. 2002.
13. Koedinger, K. R. "Cognitive tutors as modeling tool and instructional model". *Smart Machines in Education: The Coming Revolution in Educational Technology*. AAAI/MIT Press, 2001.
14. Lumet, S. *Making Movies*. Random House, Inc. 1995.
15. Mamet, D. *On Directing Film*. Penguin Books. 1991.
16. Mankoff, J. and Abowd, G.D. "Error Correction Techniques for Handwriting, Speech, and other ambiguous or error prone systems." GVU TechReport GIT-GVU-99-18. June 1999.
17. Mankoff, J., Hudson S.E., and Abowd, G.D. "Providing Integrated Toolkit-Level Support for Ambiguity in Recognition-Based Interfaces", Proc. of CHI 2000, 368-375, 2000.
18. Oviatt, S.L. "Taming Speech Recognition Errors Within a Multimodal Interface", in CACM, 43(9), 45-51, 2000.
19. Suhm, B., Myers, B.A., Waibel, A. "Multimodal error correction for speech user interfaces." TOCHI, 8(1), 60-98, 2001.
20. Yankelovich, N., Levow, G., and Marx, M. "Designing SpeechActs: Issues in Speech User Interfaces", Proc. of CHI '95, 369-376, 1995.
21. Zoltan-Ford, E. "How to Get People to Say and Type What Computers Can Understand", International Journal of Man-Machine Studies, 34 (4), 527-547, 1991.