Automated Videography for Residential Communications

Andrew F. Kurtz*, Carman Neustaedter, and Andrew C. Blose

Eastman Kodak Company, Rochester NY

ABSTRACT

The current widespread use of webcams for personal video communication over the Internet suggests that opportunities exist to develop video communications systems optimized for domestic use. We discuss both prior and existing technologies, and the results of user studies that indicate potential needs and expectations for people relative to personal video communications. In particular, users anticipate an easily used, high image quality video system, which enables multitasking communications during the course of real-world activities and provides appropriate privacy controls. To address these needs, we propose a potential approach premised on automated capture of user activity. We then describe a method that adapts cinematography principles, with a dual-camera videography system, to automatically control image capture relative to user activity, using semantic or activity-based cues to determine user position and motion. In particular, we discuss an approach to automatically manage shot framing, shot selection, and shot transitions, with respect to one or more local users engaged in real-time, unscripted events, while transmitting the resulting video to a remote viewer. The goal is to tightly frame subjects (to provide more detail), while minimizing subject loss and repeated abrupt shot framing changes in the images as perceived by a remote viewer. We also discuss some aspects of the system and related technologies that we have experimented with thus far. In summary, the method enables users to participate in interactive video-mediated communications while engaged in other activities.

Keywords: Videography, cinematography, video-mediated communications, automated capture, scene transitions, video conferencing, videophones, webcams, media space, telepresence.

1. INTRODUCTION

While the video-telephone was anticipated in the 1914 serialized novel "*Tom Swift and His Photo Telephone*", and AT&T Bell Labs demonstrated the first version of their PicturePhone™ system in 1964, the conceptual desire has largely gone unfulfilled by technological progress. The PicturePhone debuted at the 1964 New York World's Fair with much fanfare, and was improved by several generations of videophones, but never attained commercial success. Technical issues, including low resolution, lack of color imaging, poor audio-to-video synchronization, and a restricted field of view, limited the performance and appeal. It was later suggested[1, 2] that the PicturePhone failed due to lack of a social need and loss of personal privacy as much as due to technical limitations. However, the current widespread use of webcams and associated video conferencing software (e.g., Skype™, Google Talk™) for personal communications over the Internet suggests that personal and social needs that motivate domestic video communications are indeed present.

Thus far, there has been little research directed specifically at the needs people may have for residentially based video communications. Media spaces, which are nominally always-on linked video connections, represent one promising area of research. The media space concept originated at Xerox-PARC in the early 1980s as prototype systems for fostering workplace collaboration. Most resulting and subsequent research[3-6] has examined the uses and reactions to video communication in the workplace, and the extension of these findings to domestic settings is questionable given the differences in expectations and behavior that typify the work and home environments. As a first example, a media space[7] that captured and transmitted still images extracted from captured video using a motion detection filter was tested in domestic environments. While families enjoyed picture sharing, privacy was an occasional issue, which was managed by simply turning the camera towards a wall. As another example, a media space system[8] for facilitating video communications between a telecommuter and in-office colleagues was evaluated with consideration given for managing personal privacy. Privacy loss was reduced using a variety of methods, including secluded home office locations, people counting, physical controls and gesture recognition, and visual and audio feedback mechanisms. However, the system was not optimized for personal communications by the residents and did not necessarily provide adequate privacy

* andrew.kurtz@kodak.com; phone: 585-477-7390; fax: 585-722-2223

Human Vision and Electronic Imaging XV, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7527, 75271G · © 2010 SPIE-IS&T · CCC code:
0277-786X/10/$18 · doi: 10.1117/12.838941

controls for home users. More recent work[9] has suggested that media spaces are indeed extensible to the home, and with appropriate value-adding features, can provide value by increasing awareness, thereby stimulating impromptu, as well as planned, communications between family and friends.

As an alternative to the always-on media space concept, teleconferencing equipment is typically used in the office environment for purposeful communications between colleagues or business partners. This technology has advanced rapidly in recent years, using the appellation "*telepresence*" to distinguish it from the marginally adequate systems of the past. Numerous companies, including Cisco Systems (San Jose, CA), and Teleris (London, UK), offer systems targeted to corporate executives. Emphasis is directed to providing image and sound fidelity, life-size images, eye contact image capture and display[10, 11], environmental aesthetics and ergonomics, as well as seamless and secure handling of large data streams through networks. For example, improved eye contact is typically achieved by hiding a camera behind a screen or beam splitter, through which it unobtrusively peers. Single-user systems are also available to enable an executive to more efficiently work from home. It has been suggested[12] that telepresence systems are an emergent technology for home use, without necessarily considering whether the technology matches residential users' needs and expectations. However, this equipment may be overmatched for domestic use, as it emphasizes features that are not necessarily priorities for the domestic user.

While webcams and associated video conferencing software such as Apple iChat™ or Skype are available with various functionalities, including multi-way chat, and image quality enhancement features, these systems likely represent a first foray into personal video communications for the home, rather than an optimized solution. Webcams not only remain tied to a computer or television, but the associated software enables user privacy controls in an incomplete manner. As another example, premised on the "smart home" model, a system based on ubiquitous computing, using wall-imbedded cameras and an adaptive projector[13] potentially enables users to multitask and freely move about their homes while maintaining eye contact. This approach, while flexible, may be too invasive and expensive for home use.

In contrast, we have broadly considered the potential expectations and requirements for personal video communications intended for domestic environments. We have surveyed current Internet video conferencing users[14] and tested domestic usage of an always-on media space with time shift recording capabilities[9]. In this paper, we report on a domestic video communication system concept[15] that is enabled by automated videography. In particular, we conducted a focus group study to probe the needs and expectations of family members concerning the use of video conferencing systems in the home. The study suggested that people have a strong desire to multitask and perform a variety of real-world activities while using a video communication system. This implies that users may be unlikely to stay in view of a camera during the course of these multitasking activities. An adaptive video conferencing system, enabled by automated videography that actively directs image capture, is one possible solution. However, such a system must seek a balance between allowing users mobility and freedom while providing viewers an enhanced experience, with image capture being adjusted gracefully to acquire images of real-time unscripted events. Our system adapts cinematography principles for image capture, preferably using two cameras - one wide field of view (WFOV) and one narrow field of view (NFOV) - to facilitate automatic capture of a scene[15]. While we have not yet developed a fully functional automated videography system, our contribution lies in the articulation of the factors that must be considered when adapting cinematography methods to domestic video conferencing.

2. USER NEEDS FOR VIDEO COMMUNICATIONS

Research has shown that people have a strong need and desire to stay connected with remote family and friends [16-18]. This involves the need to gather an awareness of remote families' or close friends' activities, locations, and status. While recent studies[16, 18] have examined how personal communications are conducted via telephone, e-mail, instant messaging, and photo and calendar sharing, there are few studies of the personal use of video conferencing, with a study of user experiences with mobile phone devices being a notable example[19].

2.1 Current webcam-based consumer experience

As one effort to address this gap, we recently surveyed current users of Internet video conferencing technologies (such as Skype, Apple iChat, and Windows Messenger™) to understand their usage experiences, privacy requirements, and potential needs[14]. This study, which was conducted via user interviews, revealed that domestic video conferencing is used either to share conversations, which are enhanced by seeing facial expressions and body language, or to share

activities, such as enabling grandparents to observe and interact with their grandchildren. While users primarily communicated with family or close friends, video conferencing events were seldom spontaneous or impromptu, but typically followed a schedule, and were nearly always preceded with a phone, e-mail, or text communication to verify availability for video. The pre-video validation was felt necessary, as video conferencing was perceived as more intrusive than the alternative technologies. Unlike conversations, which tend to be short and purposeful, activity-sharing events often extended in duration, with the video link becoming more like a media space. Many users also showed a need for mobility, as they manually moved webcams or laptop computers about rooms or the house to redirect image capture. This study clearly showed the value of domestic video communications and a variety of needs related to managing awareness and availability, enhancing capture mobility, and enabling user privacy controls.

2.2 Experience with a domestic media space

As a parallel effort, we have also developed a prototype domestic media space, the Family Window[9], which was used and evaluated by families over the course of several weeks. As a media space, this system was always, or nearly so, capturing video content whether activity was occurring in front of the camera or not. While this system provided some privacy control features, including separate audio and digital venetian blinds, it primarily provided features to enable increased awareness and communications between participants. As one example, a "knocking" feature was provided to enable remote viewers to notify local users that they were available by emitting a knocking sound into the local environment. Additionally, users could leave each other handwritten messages using a touch screen. Video could also be recorded for time-shifted viewing, based on activity or motion sensing, if a remote viewer was not available to watch the streaming video in real time. Whether live or recorded, video was sent over the Internet at 1 fps. Testing with the Family Window prototype revealed a wide range of user responses, including increased awareness and connectedness, as the video link prompted spontaneous interactions between remotely located family members[9]. Again, users revealed a need for capture mobility, as they manually directed capture towards activities they wanted to share.

2.3 Bringing voice to user needs and expectations

As yet another effort, which is the main subject of this paper, we address the issue of user mobility with a video communications system enabled by automated videography. To obtain an anticipatory view of consumer needs and expectations, a focus group study was conducted to examine generational and stage-in-life similarities and differences in communications and technology use. Participants were split into four demographically separate focus groups: "tech savvy teens" (15-17 year olds), "technically savvy college students" (18-24 year olds), "young families" (25-34 year olds), and "older families staying in touch with remote family" (35-55 year olds). Each moderator-led session involved initial discussions of existing communications technologies, followed by more detailed discussions that probed participants' needs and expectations for domestic video conferencing, including comparisons of a conceptual video conferencing system to other related technologies (e.g., Apple iChat, Skype, etc). The sessions lasted two hours and each participant received a small gratuity.

The initial discussion period revealed a wide range of user attitudes towards existing communication technologies. When asked how long-distance communications could be improved, every group, without prompting, cited the lack of personal video communications technologies. Participants clearly viewed personal video communications as a yet unmet need, with existing products not providing the ability to see faces or read body language in real time. The participants viewed futuristic video conferencing systems as permitting hands-free operation, as they placed great importance in being able to move around and perform other activities while communicating. Participants said that multitasking would include both complex tasks (such as dinner preparation) as well as simple activities (such as walking around a room). This is akin to the way people often use telephones while performing other activities. Interestingly, participants generally did not consider currently available webcams to adequately support this real-world multitasking. This is because webcams are generally tied to a computer, typically provide poor video quality, and constrain user activities with fixed fields of view. Participants also generally desired an optimized, stylized, stand-alone device of modest size, which would be appropriate for the kitchen or living room.

The option of improved camera capabilities was then discussed with the participants, including the use of cameras with higher resolution sensors and optical zoom, as well as automated camera control, to keep the primary users in view of the camera and a remote viewer. Interestingly, the participants tended to view themselves as a person being videoed, rather than as a remote viewer. As a result, many participants associated higher image quality than is present in current webcams with a greater risk of privacy loss, and thus they desired greater privacy controls than are present with current

webcams. The participants generally indicated that a system having an audio-only option, during which the camera would be visibly blocked, was a required minimum for privacy protection. They generally desired more than that, but also viewed more technically complex privacy controls with skepticism. As a result, they also viewed systems concepts having multiple cameras, or pan tilt and zoom controls, or automatic or remote user camera controls, with wariness, even though these components can enable hands-free use by a local user as well as an enhanced viewing experience for a remote viewer. Secondary features, such as video recording or editing, Internet access, or dual use as a digital picture frame or television, received mixed responses.

In general, the respondents anticipated that this type of system would primarily be used to communicate with friends and family for personal, one-to-one conversations, lasting 15-20 minutes. The respondents considered issues of personal appearance, background appearance, or changes therein, or the use of audio only, as being potentially problematic. In general, the older families group was the most welcoming to these technologies, both in terms of perceived need as well as acceptance of greater automatic functionality or optional features. Likewise, the tech savvy teens group was also welcoming, and anticipated using the device to record content for uploading to *YouTube* or *MySpace*, whereas, the reactions of the two intermediate groups, and particularly the technically savvy college students, suggests that they may be reluctant adopters. While they value the potential to communicate with their friends, the potential for increased personal (and perceived as more formal) communications with parents was not necessarily desired.

In conclusion, while our subjects anticipated that technologies that provide enhanced video quality during spatially unconstrained user activity are ideal, they are uneasy about potential solutions. However, the respondent reactions also suggest that the different demographic groups may prefer systems equipped with different feature combinations. The focus group study also suggests that the presence of everyday video communications may motivate the evolution of new etiquettes to help manage these social interactions.

2.4 Imaging activity for viewing

The research suggests one vision for personal video communications, in which users freely engage in their activities while also interacting with remote viewers. Inherently, user activities can range from purposeful to spontaneous, involve one or several people, and be spatially constrained or expansive. Transitions can also occur, abruptly or gradually, between these activities, which are unscripted and occur in real time. Therefore, from the perspective of the remote viewer, balance can be sought between providing high-resolution close-up images of the captured subject and their activities versus the uncertainty of a subject's next actions, and the potential for subject loss as the individual leaves the captured field of view. Moreover, as the capture conditions, and particularly shot framing, are transitioned to adapt to changing user activity, it is preferable if the remotely provided video changes gracefully rather than in a jerky or strobing manner, if possible. To accomplish this, a system that adapts the principles of cinematography to the capture of real-time unscripted events is anticipated.

However, in comparison to telepresence conferencing, domestic video communications likely involves smaller displays and more informal and active behavior. Thus, the expectations for eye contact[10, 11] and the value of imbedded cameras or eye gaze corrective algorithms are likely reduced.

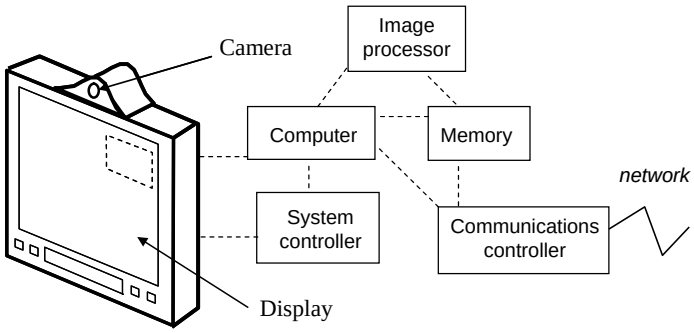

Figure 1. System configuration.

3. SYSTEM CONCEPT OVERVIEW

As shown in Figure 1, the automated videography-based system comprises one or more cameras, microphones, speakers, displays with split screen capability, and an internal computer supporting multiple specialized algorithms, including image processing that enables a user to see and hear, and thus communicate with, a person at a remote location. The system[15] would support a user interface for controlling display, camera, and audio settings, as well as privacy, call placement, recording, and automation settings. An audio subsystem, including speakers, microphones (possibly directional), noise reduction, and echo cancellation is also included.

As the system acquires images of people involved in real-time unscripted events, the use of multiple cameras, and particularly a WFOV camera that observes a large area, and a NFOV camera that observes a lesser area within the WFOV, can be particularly useful. This is shown in Figure 2, where the NFOV camera has automatic pan, tilt, and zoom (PTZ) adjustment capabilities, such that the captured NFOV area can be changed in size and location in response to signals from either a local or remote system. While digital zoom of a captured WFOV image can be used, an optically zooming NFOV camera will likely provide better image quality. A camera can also be provided behind the display screen to enable eye contact image capture[10, 11].

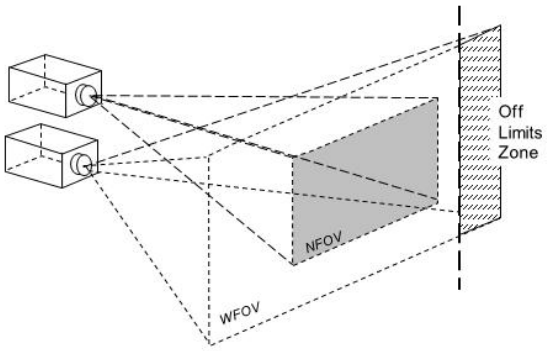

Figure 2. Two-camera setup and privacy zones.

The imbedded computer, which not only provides image processing but a user interface including a privacy interface, and a contextual interface for interpreting scene content and determining system responses, is an intelligent agent comprising a series of algorithms. It includes call management algorithms to process identity, address, encryption, and data transfer protocol information across a network. It also can include algorithms for face detection, body shape detection, motion detection, the counting and tracking of people, and quantifying user activity, which can operate in real time, or nearly so, at least on a burst basis. However, the key to real-time video capture of unscripted events depends on the interpretation of contextual cues regarding user activity and the automated response thereto.

3.1 System workflow

A simplified illustration of system operation is depicted in Figure 3 using the formalism of a phone call. When a call between local and remote users commences, an initial audio-only transmission can be followed by video transmission, using default or user-defined settings that can affect video or audio capture, privacy, recording, capture mode selection (such as lock and follow) and other attributes. Then, as a local user participates in the communication event, the system would monitor the user activity using a transition test, to determine whether any changes in user activity are consistent with the current video capture settings (*an intra-scene transition*) or are indicative of changes to new video capture settings (*an inter-scene transition*). In the former case, scene capture management algorithms can determine and apply small changes in video settings (for example, focus, zoom, pan, tilt, framing, or brightness) that are consistent with the current shot framing. In the latter case, where changes in user activities push beyond the current framing, a transition process is initiated to determine changes in shot framing and other video capture settings, and then applies them in a graceful way if possible. Once new conditions are established, the process of transition testing and scene capture management or transition processing continues until video transmission is concluded or re-defined by the users. Similar, but simplified, workflows can manage automatic capture for other types of systems, such as media spaces.

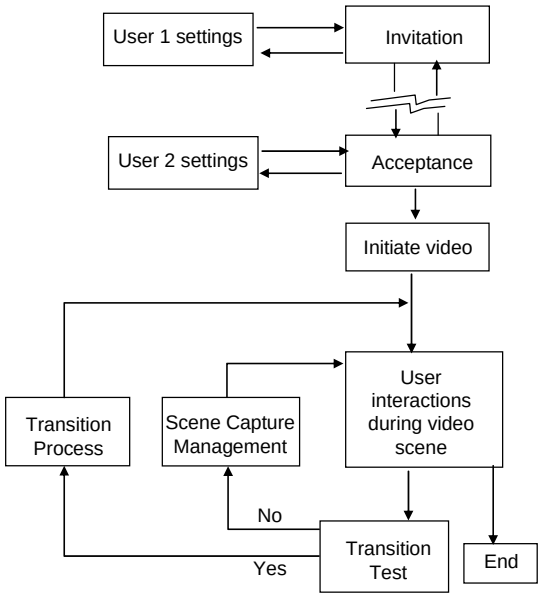

Figure 3. System workflow.

3.2 Types of communication events

While communication events between local and remote users will often involve just a few individuals who are engaged in focused conversations, a wide range of possibilities exist. These events can be purposeful or spontaneous, involve single or multiple individuals, or be motivated by special events (such as a birthday party) or combinations thereof. As user expectations can vary for different types of events, the automated video capture settings, relative to field of view, privacy, shot selection and framing, shot duration and transition rates, and other factors, can likewise change.

Thus it is useful to define automatic capture with various user-selected capture modes, including fully automatic, lock and follow, hierarchical, or those based on event classifications (with responses linked to contextual or activity cues).

System function is better understood by example, as illustrated in Figure 4, which depicts a simplified two-scene communication event. In the first scene, a local user sits at a table, looking at a display where an image of a remote user is shown. If the two users are engaged in a purposeful conversation, then the imaged field of view may be narrow. Then as the communication event continues, the users may relax and transition to more spontaneous activity, where a wider image field of view may be more appropriate. In Figure 4, this evolution is suggested by the first scene (A1) with the seated local user, followed by the second scene (A2) where the local user now stands behind the chair.

3.3 System response

In turn, this requires a system that supports video capture of real-time unscripted events with a variable number of participants and operating conditions, which implies ongoing interpretive environmental sensing and capture management automation. As an approach, the analysis algorithms can assess the available cues to detect both small and large changes in user activities that warrant changes in the video capture, and the system can change capture conditions accordingly. In that regard, it is a useful construct to define *an intra-scene transition* as relating to capture setting changes that are consistent with the camera shot of a current scene, and to define an *inter-scene transition* as relating to capture setting changes that occur during transitions between scenes.

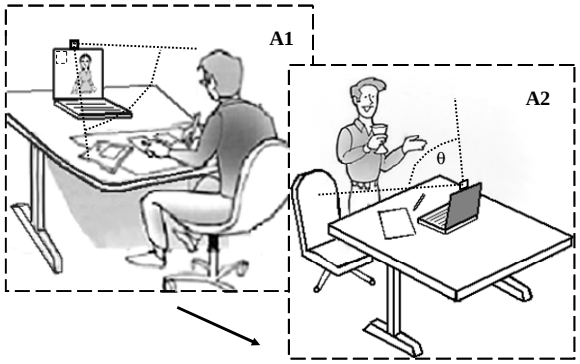

Figure 4. Shot framing and transitions.

As an automated system, decisions on shot framing and shot transitions are determined by interpreting current and previous video content. This video content, as an expression of user activity, can be interpreted using semantic cues or cues based on the location and duration of user activity. While semantic cues, involving knowledge of user classifications, user identity, event classifications, user behaviors, and other factors, can be used, they require the system to acquire and interpret complex social information. By comparison, automated videography based on quantitative measurements and assessments of user activity, relative to shot framing, is a more versatile approach that can adapt to transitions between spontaneous, purposeful, single subject, and multiple subject events in any combination. Semantic information can then affect how the system responds to measured changes in the activity-based cues, and whether they are deemed consistent with an *intra-scene transition* or *inter-scene transition* with respect to the current shot framing. The user-selected video capture modes then can constrain the automated response.

3.4 Enabling user privacy

It can be expected that privacy concerns will influence user acceptance and feature requirements[8, 9, 14, 20]. While audio-first and visible camera blocking are useful, they are likely insufficient mechanisms. For example, local users may want to control video transmission to a remote viewer, including limiting video recording or local camera control options by the remote viewer. As another privacy control, users may want to limit the field of view that can be captured or transmitted by the cameras. While constraints can be provided environmentally (by controlling lighting, closing doors, etc.), the field of view can also be limited electronically. As shown in Figure 2, a portion of the field of view captured by the WFOV camera can be defined as off-limits for spatial privacy. Image content from that area could then be cropped or shaded prior to transmission. Likewise, image capture by the NFOV camera, relative to the PTZ controls and image cropping, can be constrained by defined off-limit areas. The local users can then use a picture-in-picture or split screen image, to have a direct visual presentation of the captured or transmitted images, which may be different, depending on applied privacy constraints.

4. CINEMATOGRAPHY AND RELATED WORK

In this case, the goal of automated videography of real-time unscripted events is to direct image capture for the benefit of the remote viewer, given the privacy constraints applied by local users, as well as the uncertainty of local user actions. As one aspect, a method is needed to automatically frame users within a camera's field of view. Some webcams

have user tracking, such as the Logitech Orbit, which directs a motorized rotating camera to follow a person's face, or the Logitech QuickCam software, which crops video to only include regions containing a face. However, more is likely required for video scenes that involve multiple users or widely variable activity transitions.

4.1 Traditional cinematography

As one approach, automated videography could borrow capture and framing conventions of cinematography[21, 22], which are quite successful for framing scripted events. Many cinematographers use a shot-framing selection convention with a range of defined shots. Figure 5 shows some sample shots: close-up (CU), medium CU, medium, medium-wide, and wide (or full). For instance, a medium shot shows the person from the waist up and is typically capable of showing two people up close in the same view (a "two-shot"); a wide shot can be wide enough to show 4–6 people in the same view. Cinematographers also often use the rule of thirds (see Figure 6), where a subject is framed off-center (along a line at the right or left third of the frame) to improve aesthetics. There are also standard transition shots (e.g., establishing shots, straight cuts, dissolves), changes in camera orientation (answer or reaction shots) and camera placements (such as the action-axis rule) to guide shot transitions and selections. While these guidelines can vary with cinematic style (e.g., dramatic, comic), the cinematographer typically films scripted events that often can be reenacted to attain the desired aesthetic look. In addition, a cinematographer has the luxury of using and moving one or more strategically placed cameras about the capture area. In short, the cinematographer has control over shot selection, framing, angular orientation, focus, timing, and point of view. By comparison, in a home, a subject's next actions are unknown to the camera, and cameras are limited in location, number, and mobility. Although cinematic image framing does not seem to translate directly to real-time videography under these conditions, the techniques suggest useful framing definitions and responses.

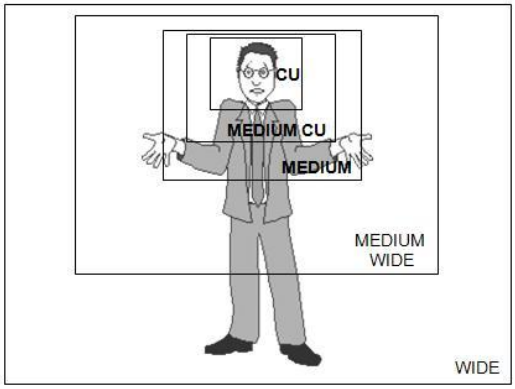

Figure 5. Cinematographic shot selection.

4.2 Automated cinematography or videography

There have been a few attempts to adapt cinematographic principles to the video capture of unscripted live events. Doubek et al.[23] provide one notable example, describing a system for capturing real-time events using cinematic principles to enhance aesthetic value. Multiple low-end networked cameras capture video from an environment based on algorithmic decisions. Notably, the cameras are placed by cinematic conventions, such as the action axis rule, and shots are selected by cinematic best-shot criteria. However, framing and re-framing relative to subject motion and for multiple subjects is underdeveloped. As another example, Pinhanez et al.[24] describe an intelligent robotic camera system that provides automatic camera framing of subjects and objects in a TV studio upon verbal request from a TV director. The goal is to operate the camera automatically without the aid of a cameraman, while changing camera pan, tilt, and zoom. Notably, this system benefits from the relatively controlled environment of the TV studio, as well as the judgment and experience of the TV director.

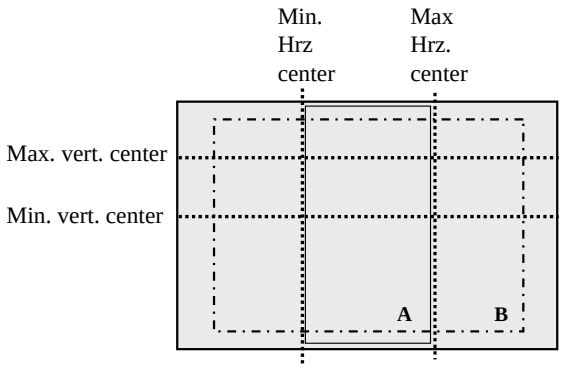

Fig. 6. The rule of thirds applied to single-user framing.

As another approach, Kim et al.[25] seek to cinematize real-time video of unscripted events. They describe a system in which a living space is populated with a multitude of cameras, including ceiling-mounted cameras and an omnidirectional camera. Each camera then captures video of ensuing events, with synchronizing time code data. The resulting video from each camera is then analyzed by algorithms using cinematographic guidelines regarding shot selection, shot perspective, zooming, panning, indecisive cuts, and the action axis rule, to classify the available shots, as well as potential shots synthesized for a virtual camera. The virtual camera shots are rendered using video from the various cameras as appropriate. The users (the director) then select the preferred shots and shot compositions to compose and post process an aesthetically-pleasing movie that progresses from scene to scene, using video from real or virtual cameras. While this method is interesting, it seems inappropriate for real-time video communications.

There are also numerous examples in the literature[26-28] where cinematography is adapted to the capture of activities of virtual actors or avatars that function within virtual worlds. While the unfolding events can be unscripted and occur in real time, cinematography in virtual metaverses is not bound by real-world restrictions. For example, Bares and Lester [26] assume an infinite number of camera angles and capture points are possible, while He et al.[27] provide camera modules that can not only modify camera capture settings and positions, but which can also change the positions and poses of the virtual actor participants.

5. AUTOMATED VIDEOGRAPHY OF UNSCRIPTED EVENTS

The present concept for automated videography emphasizes shot selection, framing, and transition management during capture of real-time unscripted events under constrained conditions. As a result, different conventions for shot selection, shot framing, and shot transitions are appropriate, which balance a subject's image size and position with the uncertainty of a subject's next action. The proposed use of *digital cinematography* should provide a pleasurable viewing experience for the remote viewer while avoiding subject loss, thereby balancing content presence with image aesthetics.

5.1 Shot selection and framing

To begin, the shot selection range is reduced, with wide, medium-wide, and medium shots being used frequently, and close-up and long shots (wider than wide) used less frequently. A wide shot may accommodate small groups of 2-6 people, while a medium-wide shot accommodates only 2-3 people, and a medium shot may be used for one person, or two people in close proximity. As a quantitative cue, a facial or head area can be monitored as a percentage of the frame area. For example, for a medium shot, a facial ROI (region of interest) can be ~ 4% of the frame area.

Unlike the classical rule of thirds found in cinematography, shot framing for video conferencing must balance the user's size in the image against the uncertainty of the user's next action. Figure 6 shows the framing zones used in our method. In the classical rule of thirds, a subject would be framed along either the "Min. Hrz center" (Minimum Horizontal Center) or the "Max. Hrz. center" (Maximum Horizontal Center) framing line. In this case, the goal is to frame a subject within the center third of the image (region A), with a facial ROI preferably crossing the maximum vertical center-line and between the minimum and maximum horizontal center-lines. As subjects can move freely within this area, without reframing being needed, aside from possible focus adjustments, the system response is dampened. However, when a subject moves outside the horizontal center-lines, then reframing can become necessary. If subject motion is modest, intra-scene transition reframing can occur with small PTZ adjustments while generally maintaining the same shot selection. Yet, if subject motion is large, an inter-scene transition to a new shot selection can result. Figure 6 also depicts a larger area (region B), an action safe zone, in which a subject can move, where the current shot selection can be preserved by a pan, tilt, or cropping positional change.

5.2 Interpretive cues for user activity

Reframing can be triggered by semantic interpretation of image data, for example, using event or people classifications. However, use of direct and indirect measures of user activity, and associated thresholds, can be very useful. For example, people counting can provide a direct input that limits framing selection. Table 1 provides a partial list of shot-framing metrics that includes both user activity and system response measures.

The activity level of local users can either increase or decrease (settle), thereby causing or allowing reframing. User activity can be measured directly using a subject movement factor, which can be a normalized product of subject movement that factors in the duration, velocity, frequency, and magnitude or movement area, relative to the captured FOV. A second measure, which is indirect and comparatively dampened, is the PTZ frequency (f_{PTZ}), or rate at which reframing occurs. For example, a modest movement of a user outside the current center framing, as measured by the subject movement factor, may be compensated for by reframing image capture to re-center the subject while maintaining the framing size. By comparison, a large movement can easily move a subject outside the current framing, necessitating reframing with a wider shot that can capture the subject, their current location, and perhaps their prior location.

Sudden subject movements can be problematic, as subject image capture can be at least temporarily lost. Thus it can be useful to measure subject movement frequencies, even when the subject is staying within current framing, to anticipate the potential for subject content loss. While the frequency of subject movement can be tracked directly, for example, by tracking head or limb movement, tracking PTZ frequency can provide a damped but effective result. For

example, while an increasing level of user activity can be captured by current framing, increases in PTZ frequency (f_{PTZ}) can indicate that the probabilities of subject loss or viewer annoyance with repeated reframing are increasing. This can then spur reframing to a wider shot. Likewise, decreasing measures of subject activity, using a subject movement factor or a frequency measure, can indicate a local user is settling. The system can then determine that an opportunity for a tighter shot exists, determine a new shot using inter-scene transition rules, and apply the new shot selection, using PTZ or image cropping as appropriate.

5.3 Shot change timing

Shot reframing can be subject to both defined shot transition times and shot transition delay times, such that when a shot change is triggered, a transition time elapses after the transition delay time has elapsed. Shot transition times define the rate at which capture settings are changed, whereas shot change delay times dampen the system response, reducing the rate of shot transitions, or reframing, and thus potential viewer annoyance.

Table 1. System and Shot Framing Metrics

Parameter	***Definition***
Sizing: ROI / image area	Facial ROI: % (face/total) image for a subject
Subject Movement Factor	a normalized product of subject movement relative to the FOV
PTZ Frequency (f_{PTZ})	rate of reframing (zooming, panning, tilting, cropping) over time
Re-center & Resize Time Delays (ΔT_d)	Delay after a change in subject activity or position before reframing
Shot Change Transition Time (ΔT_{ST})	time (frames) for transitioning from current capture settings or shot to new settings or a new shot
Scene change probability (P_{IA} or P_{IE})	Probability to change capture settings for ***intra-*** or ***inter-scene*** transitions
Shot selection probability (P_{SF})	probability in determining the next shot

Shot reframing can occur over a defined scene change transition time (or shot change transition time) ΔT_{ST}, which is an allotted time for transitioning from current video capture settings (including shot framing) to new settings. This transition time ΔT_{ST} depends upon the current shot framing, the new shot framing, and the allowed camera slew or zoom rates. Transition timing for small changes in subject activity (intra-scene) can be more casual than transition timing for large changes in subject activity (inter-scene). This is because users are likely to perform more small movements (intra-scene changes) than large movements (inter-scene changes), and transitioning the video quickly in response to small movements would create video jitter. Image cropping changes or camera selection can be instantaneous, but gradual changes are preferred to provide a better real-time viewing experience.

Shot change delay times (ΔT_d) are applicable whether the amount of user activity increases or decreases. As with the shot transition time ΔT_{ST}, different delay times can be used for shot framing changes for intra-scene transitions than for inter-scene shot changes, as well as for different shot change combinations. Delay times can become longer as the shots become progressively tighter. For example, the delay time before transitioning from a wide shot to a medium-wide shot of a single subject may be ~ 40 seconds, while the delay time for transitioning from a medium-wide shot to a medium shot may be ~ 80 seconds. These long delay times help balance the risk that the subject has not settled with the goal to provide more detailed images to a viewer. To minimize subject loss, the delay times for transitioning from a tight shot to a wider shot can be more rapid (a few frames).

However, once in a relatively tight shot, the delay times (ΔT_d) for responding to increased subject movement, whether measured by a subject movement factor or a PTZ frequency (f_{PTZ}), are much reduced. For example, when a subject moves laterally relative to the current framing, outside the horizontal center lines but still within the frame, the system may wait several seconds or longer to see if the subject returns to frame center before reframing the subject with pan or crop adjustments. On the other hand, if the subject moves rapidly beyond the edge of the frame, a time delay before reframing to a wider shot may only be a few frames or less. Of course, a local subject may move out of the current framing and reframing or new shot selection is prevented by the privacy limits provided for that event by a local user.

5.4 Single vs. multiple subjects

Scene transitions and shot reframing are somewhat different when multiple users are present or enter a scene. Here the goal is to locate all users within a wider region A (the central ~ 60% of the image) than in Figure 6. As with a single user, a few people can settle into a FOV, thereby enabling a tighter shot after an appropriate settling time has passed. In the case that a user leaves the central area but remains within a current frame, reframing can occur. Likewise, reframing

to a wider frame can occur if a user leaves the current frame entirely. Shot transition delay times again apply, but as the risk of subject loss is greater, they tend to be shorter than with single subjects. However, the use of shot transition delay times provides margin to determine if the subject leaves the entire allowed capturable FOV first, thereby preventing a low value-added shot transition. Shot-reframing decisions can also be linked to user identities or classifications.

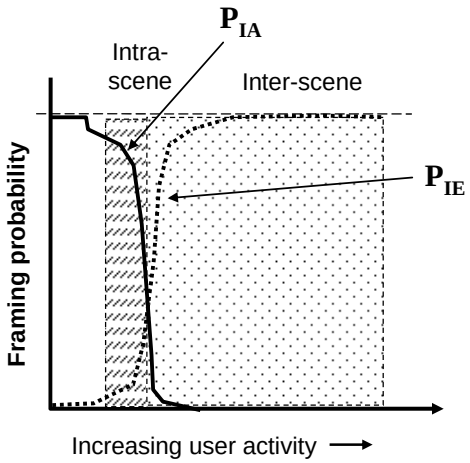

Figure 7. Shot transition probabilities – for a single user captured in a medium shot.

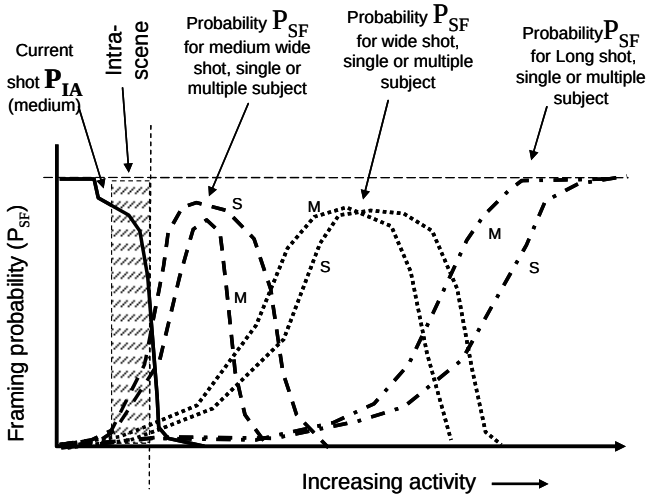

Figure 8. Probability for scene transitions and shot selections.

5.5 Shot selection and re-framing probabilities

As stated previously, single subjects can be captured in a wide range of shots, spanning from long shots to close-ups. However, as the number of subjects increases, the use of the tighter shots diminishes. Stated another way, the probability of the system being in a given shot depends on the number of subjects. Likewise, as subject activity changes, the probability of staying in a current shot and reframing with the current shot (P_{IA}) for an intra-scene transition, reframing to a new shot (P_{IE}) for an inter-scene transition, and shot selection (P_{SF}) for the new shot, can all be considered relative to increasing levels of user activity.

Figure 7 gives a representational example of the probabilities relative to staying in a current (medium) shot or changing to a new shot, relative to increasing user activity. As can be expected, the tighter a current shot is, the less likely it is that increasing levels of user activity can be supported by the current shot. While it can be relatively easy to determine that an inter-scene transition is occurring, the determination of the best next shot can be more difficult. These probabilities P_{SF} (see Figure 8) will also be different for a single subject, a small number of subjects, or a large number of subjects, or when activity transitions change the relative number of subjects. Therefore, it can be valuable to calculate confidence values for shot selection based on subject activity (using the subject movement factor or PTZ frequency), the number of subjects involved, and the current shot selection. These probabilities can then be used to aid the system to determine the next shot during an inter-scene transition. It can also be valuable to track the evolution in subject movement, PTZ frequency, the number of subjects, and shot selection over a number of scene transitions in communication events, to develop a statistical history with respect to activity levels and scene transitions.

Shot selection is not limited to the standard shots shown in Figure 5, as intermediate shots (e.g., between medium and medium wide) can occur if the subject movement (area, magnitude, frequency, and direction) does not match well with a standard shot. In a sense, each shot, as well as intermediate shots, can be defined and selected according to associated facial or body ROIs, an allowed number of subjects, an appropriate range of subject movement, or other factors. An intermediate shot can also occur on an interim basis if the system is uncertain of the best shot selection.

6. SYSTEMS APPROACH

The software system required to implement automated videography in real time is complex, necessitating a software architecture, as shown in Figure 9, that decomposes the activities and responsibilities into separate subsystems. The

environmental sensing subsystem is responsible for processing "raw" input from the audio/video capture devices and converting this input into sense events. Examples include: face detections, subject identities, and object detections. Sensed events are delivered to the environmental monitoring subsystem which is responsible for maintaining state about the environment. Sensed events are processed to maintain and update the state of the environmental tracking of the identity and location of subjects and objects. The intelligent call agents are responsible for acting on changes to the environment delivered by the environmental monitoring subsystem. One or more agents would be responsible for implementing automated cinematography by interpreting and reacting to environmental changes. Actions required by the cinematography agent are carried out by the camera control subsystem.

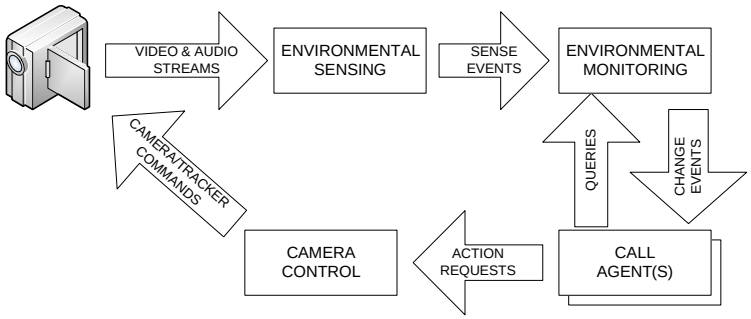

Figure 9. Software architecture.

7. CONCEPT DEVELOPMENT AND CONCLUSIONS

The authors have described a residentially targeted video communications system[15] that uses automated videography to enable local users to freely multitask while conversing, or alternately engage in purposeful conversations, while simultaneously providing a pleasurable viewing experience for remote participants. Thus far, a working prototype of the described system has not been developed in entirety. Rather, a variety of exploratory probes has been developed, including an environmental sensing module, and a basic two-way, real-time, Open CV based prototype version of this system with face detection and call management interfaces, as well as a preliminary spatial privacy interface. As another interim step, we have experimented with the Family Window time-shifting media space[9], which allows us to explore interactions in the domestic space, without the complications of active automated camera control. Despite a lack of evaluation and complete system development, our work still contributes a method for adapting cinematic framing to domestic video conferencing or other applications. We have also outlined various challenges to creating smooth video transitions and visually pleasing frame compositions, along with features to address them.

8. ACKNOWLEDGEMENTS

We wish to specifically acknowledge the efforts of Kathleen Costello for arranging and conducting the consumer focus group activity, and Tejinder Judge for her survey of webcam video conferencing users[14]. We are also grateful to Phoury Lei and David Barnum for their development help with both the Family Window[9] and this system.

9. REFERENCES

[1] Noll, A., "*Anatomy of a Failure: Picturephone Revisited,*" Telecomm. Policy, Vol. 16, pp. 307-331, (1992).

[2] Lipartito, K., "*Picturephone and the Information Age: The Social Meaning of Failure,*" Technology & Culture, Vol. 44, pp. 50-81, (2003).

[3] Fish, R., Kraut, R., and Chalfonte, B., "*The VideoWindow System in Informal Communications,*" Proc. CSCW 1990, pp. 1-11, ACM Press, (1990).

[4] Bly, S., Harrison, S., and Irwin, S., "*Media Spaces: Bringing Together a Video, Audio and Computing Environment,*" Communications of the ACM, Vol. 36, pp. 28-47, (1993).

[5] Harrison, S., Bly, S., Anderson, S., and Minneman, S. (1997). "*The Media Space,*" in K. Finn, A. Sellen, and S. Wilbur (eds.) Video-Mediated Communication, Lawrence Erlbaum, Mahwah, NJ, pp. 273–300, (1997).

[6] Coutaz, J., Bérard, F., Carraux, E., and Crowley, J., "*Early Experience with the Mediaspace CoMedi,*" Proc. of EHCI 98, pp. 57-72, (1998).

[7] Conversy, S., Mackay, W., Beaudouin-Lafon, M., and Roussel, N., "*VideoProbe: Sharing Pictures of Everyday Life,*" Proc. IHM 2003, ACM Press, (2003).

[8] Neustaedter, C., and Greenberg, S., "*The Design of a Context-aware Home Media Space for Balancing Privacy and Awareness*," Proc. of Ubicomp2003, pp. 297-314, (2003).

[9] Judge, T. K., Neustaedter, C., and Kurtz, A. F., "*The Family Window: The Design and Evaluation of a Domestic Media Space*," to be published in Proc. of CHI, 2010.

[10] Gale, C., and Monk, A., "*Where Am I Looking? The Accuracy of Video-mediated Gaze Awareness*," Perception & Psychophysics, Vol. 62, pp. 586-595, (2000).

[11] Vertegaal, R., Weevers, I., Sohn, C., and Cheung, C., "*GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Direction*," CHI 2003, ACM Press, pp. 521-528, (2003).

[12] Cringley, R. X., "*The Next Killer App: Telepresence May Come to Your House Next Year*," published online at The Pulpit, www.pbs.org, PBS (2007).

[13] Tapia, E., Intille, S., Rebula, J., and Stoddard, S., "*Concept and Partial Prototype Video: Ubiquitous Video Communication with the Perception of Eye Contact*," Ubicomp 2003, (2003).

[14] Judge, T. K., and Neustaedter, C., "*Sharing Conversation and Life: Video Conferencing in the Home*," to be published in Proc. of CHI, 2010.

[15] Kurtz, A. F., Border, J. N., Costello, K. M., and Parada, Jr., R. J., "*A Residential Video Communication System*," US Patent Publication 20080298571.

[16] Neustaedter, C., Elliot, K., and Greenberg, S., "*Interpersonal Awareness in the Domestic Realm*," Proc. OzCHI 2006, ACM Press, pp. 15-22, (2006).

[17] Romero, N., Markopoulos, P., van Baren, J., de Ruyter, B., Jsselsteijn, W., and Farshchian, B., "*Connecting the Family with Awareness Systems*," Personal and Ubiquitous Computing, Vol. 11, pp. 299-312, (2006).

[18] Tee, K., Brush, A. J. B., and Inkpen, K. M., "*Exploring Communication and Sharing between Extended Families*," International Journal of Human-Computer Studies, Vol. 67, pp. 128-138, (2009).

[19] O'Hara, K., Black, A., and Lipson, M., "*Media Spaces and Mobile Video Telephony*," in Media Space: 20+ Years of Mediated Life, Springer, pp. 303-324, (2009).

[20] Boyle, M., Neustaedter, C., and Greenberg, S., "*Privacy Factors in Video-based Media Spaces*," in Media Space: 20+ Years of Mediated Life, Springer, pp. 97-122, (2009).

[21] Mascelli, J. V., "*The Five C's of Cinematography: Motion Picture Filming Techniques*," Silman-James Press, Los Angeles, (1965).

[22] Brown, B., "*Cinematography: Theory and Practice*," Focal Press, (2002).

[23] Doubek, P., Geys, I., Svoboda, T., and Van Gool, L., "*Cinematographic Rules Applied to a Camera Network*," Proc. of OmniVis 2004, pp. 155-230, (2004).

[24] Pinhanez, C. S., and Bobick, A. F., "*Intelligent Studios: Using Computer Vision to Control TV Cameras*," ICJAI '95 Workshop, pp. 69-76, (1995).

[25] Kim, H., Sakamoto, R., Kitahara, I., and Kogure, K., "*Cinematized Reality: Cinematographic 3D Video System for Daily Life Using Multiple Outer/Inner Cameras*," IEEE CVPRW'06, pp. 168-175, (2006).

[26] Bares, W., and Lester, J., "*Cinematographic User Models for Automated Realtime Camera Control in Dynamic 3D Environments*," Proc. of User Modeling, pp. 215-226, (1997).

[27] He, L. W., Cohen, M. F., Salesin, D. H., "*The Virtual Cinematographer: A Paradigm for Automatic Real-Time Camera Control and Directing*," SIGGRAPH '96, pp. 217-224, (1996).

[28] Tomlinson, B., Blumberg, B., and Nain, D., "*Expressive Autonomous Cinematography for Interactive Virtual Environments*," Fourth International Conference on Autonomous Agents, Agents 2000, pp. 317-324, (2000).