Teleconferencing Using an Attentive Camera System*

Volker Krüger[1], Rainer Herpers[1,2], Kostas Daniilidis[1,3], and Gerald Sommer[1]
[1] Computer Science Institute, Christian-Albrechts University Kiel
Preußerstr. 1-9, 24105 Kiel, Germany
Tel: ++49-431-560496, FAX: ++49-431-560481
email: [vok,rhe]@informatik.uni-kiel.de
[2] University of Toronto, Department of Computer Science
Toronto, Ontario, Canada
[3] GRASP Laboratory, University of Pennsylvania,
Philadelphia, USA

Abstract

State-of-the-art teleconference systems are not capable of recording a discussion of an entire group. The reason is that limited bandwidth of computer networks only allows for the use of cameras with a limited resolution just large enough to transmit a detailed image of a single person. In this contribution an attentional camera system for teleconferencing is presented that is able to follow a conversation of a group of people by actively controlling the camera. The system acts like a silent observer who continuously gazes at the person who is talking. Detection and tracking is based on visual as well as on acoustic cues. The system is designed as a Perception-Action Cycle (PAC) which is an example of a behaviour-based vision system.

1 Introduction

Teleconferencing systems allow remote users to get in contact with each other not only acoustically but also visually. Even if users are hundreds of miles apart, they would be able to talk with each other as if they were in the same room. This allows them to emphasize their talking with gestures and facial expressions. E. g. users could even see each other blush or turn pale. Another important aspect of teleconferencing systems is that more than one person at one place is able to communicate instantaneously with other groups at different locations. Furthermore, it could save in traveling costs, jet lag adaptation and, above all, time. These aspects may become increasingly important in the future. This would be especially true for international companies, which have to coordinate their global policies and economical strategies, for example.

However, due to limited bandwidth, current teleconferencing systems use low resolution cameras which are generally fixed or positioned on a particular location with a given focal length. Such systems are either single user systems or overview systems which are unsatisfactory in their resolution. In this contribution we address this problem and present an attentive camera system (ACS) which may replace teleconferencing systems with stable cameras. As input an ACS uses a RGB-camera system with three controllable degrees of freedom (pan, tilt, zoom) and a directional microphone. The microphone is mounted on the camera such that its "acoustic axis" (the sensitive section) is parallel to the visual axis of the camera. If the camera is oriented such that the person's head is located in its visual field, this allows for a verification of a talking person since the microphone also receives a related acoustic signal. Based on this system design, the ACS is able to follow a discussion or a conversation of a group at one location similar to a silent observer. In general, the ACS is intended to act like a virtual camera man who focuses on the person who is speaking.

In a typical scenario of a discussion or conversation, one person is talking while all other participants are listening. The speaker talks which means that his/her voice can be heard and his/her motion and gestures can be seen. On the other hand, listeners primarily do not talk[1] and may show considerably less motion than the speaker. Therefore, motion, gesture and voice are attentive cues that might clearly identify a speaker. These cues are exploited by the ACS proposed here. The ACS detects the visual attentive cues to select a person that is talking. Then the ACS verifies if a particular candidate is talking by first orienting the camera appropriately towards him/her in order to exploit the acoustic cues. If no acoustic cues can be detected, the ACS searches for another speaker. If the acoustic

*This work is partially supported by the DFG, Grands: So 320/1-2, Ei 322/1-2 and He 2967/1-1.

[1] as a matter of politeness

cues are proved successfully to originate from the focussed speaker (s)he is zoomed in and tracked while (s)he is talking. During the continuous tracking of the face which is zoomed, the system is able to detect additional and more detailed facial features[2] [3]. It should be mentioned that the directional microphone is not used to record the acoustic signals. It is only used to verify the acoustic cues.

An approach for an active teleconferencing system has been presented in [2]. To localize a person, colour and eye blink detection is applied. Tracking is established using a correlation technique. However, no integration of the presented techniques into a single system has been done. In [13], the aspect of limited network bandwidth is tackled by a teleconferencing system that defined a region of interest within the camera image that is transmitted with high quality while the rest of the image was transmitted with less quality. In their system the selection of the region of interest is performed manually and the camera was fixed. A virtual space teleconferencing system is presented in [8] that gave their users the sensation of being all at the same site. Their system uses two large screens to provide a real-time reproduction of 3D whole body human images. The system needs data gloves, tape marks, and magnetic sensors to achieve real–time reproduction, which can not be considered to be natural human communication. The system presented in [7] uses a camera that is able to consider a field of view of 360°. It is able to view an entire conference table, when the camera is positioned in the center of the table. However, the system is not able to zoom in on the speaker.

The general methodology of our ACS is introduced in detail in the next section. In section 3 we briefly present some details concerning implementation aspects and experimental results. We close up our contribution with some concluding remarks.

2 Methodology

The intended purpose of the ACS is to record a discussion of N people possibly positioned around a table in front of a camera. In order to record the entire scene containing all participants the camera is initially set to wide angle with a default camera orientation. When one person starts talking, the ACS performs the following actions: Firstly, it detects the person which is talking. Secondly, it directs the camera to that person and focuses him/her. Finally, if all evaluation checks have been successfully computed, it zooms in on the speaker's head. During the computation, visual as well as acoustic attentive cues are evaluated[3]. The underlying assumption is that the motion, facial gestures, and acoustic cues would identify a speaker reliably. "Focusing" is defined here by directing the pan and tilt of the camera to the speaking person in particular to the person's head. "Zooming" refers to focusing and zooming in so that the camera records a head-shoulder view of the speaker (see fig. 1, right). The zooming action also involves a continuous tracking of the face which may be moving.

Figure 1: Typical scenes recorded by the ACS: Left, a scene with two people are shown. Each head is relatively small. However, when a person starts talking, (s)he is zoomed in (right) so that a remote observer is able to easily understand the person's facial expressions and gestures.

In order to detect, focus, and zoom the speaker, the ACS evaluates the visual attentive cues and sorts these cues in decreasing order with respect to their saliency. The ACS focuses first that region which attracts the most attention, in other words, the region from which the most and highest attentive cues come. While the speaker is zoomed in, the ACS keeps the head of the speaker approximately in the center of the camera image by continuously correcting the camera values. If the speaker stops talking, much less visual and acoustic attentive cues would be detectable. The ACS can detect these changes and returns to the wide angle view mode. In other words, the ACS is reset to the initial orientation and the scene is zoomed out. Subsequently, the scene is inspected again to detect and focus a new speaker. In a situation where no one speaks, the ACS remains in the wide angle view position until remarkable motion and acoustic cues can be detected. This system behaviour corresponds well with natural human behaviour which also directs attention to areas of increased motion and acoustic features.

An example of this processing can be seen in fig. 1. The left image shows a typical scene recorded in wide angle view. This view is similar to the view of today's teleconferencing systems. Facial expressions can hardly be recognized. The image to the right shows

[2] This knowledge about the content of the image may be used to support an efficient coding of the recorded video sequences.

[3] In the following, the term "attentive cues" will always refer to visual and acoustic attentive cues of a speaker such as gesture, motion and voice.

the camera view of one of the persons when (s)he is zoomed in. It can be seen in this sample head-shoulder image that now facial expressions can be recognized more clearly.

To be precise the system is organized as an Perception-Action-Cycle (PAC) which is decomposed into three different layers (fig. 2, right): the *perception-layer*, the *evaluation-layer* and the *action-layer*.

The perception-layer supplies a set of possible perception methodologies that allows the system to exploit different attentive cues. The different types of attentive cues used by the ACS are motion, colour, and acoustic cues.

The evaluation layer is used to evaluate the perceived information and to instantiate an appropriate action of the ACS.

The action layer establishes the control of the camera. It executes and controls the camera actions.

The evaluation layer is the most essential layer because of two reasons: First, the evaluation layer selects and evaluates the attentive cues derived by the perception layer. What type of attentive cues are used and how they are evaluated at each time step depends on the internal state of the camera as well as on the situation in the scene. Second, the evaluation layer instantiates the camera action needed to perceive new and different attentive cues.

The concepts used for the design of the evaluation layer is inspired by attentive approaches published in [4, 10]. The evaluation layer has three different attentive states $S = \{\texttt{base}, \texttt{focus}, \texttt{zoom}\}$. The system starts in a wide-angle mode called `base` mode and evaluates motion and colour cues in order to direct its attention to a limited part of the scene. When the evaluation layer has detected and evaluated appropriate visual cues it subsequently initiates a change of the system state to the `focus` mode and the action layer directs the orientation of the camera appropriately such that the object is in the center of gaze of the camera. In `focus` mode the evaluation layer ensures a tracking of the focused object which is achieved here by exploiting colour cues only. Focusing the object allows to limit the computational resources of the ACS to a single chosen object. The task of the `focus` mode is to verify the acoustic cues of the tracked person. For this, the tracking of the person object ensures a correct orientation of the directional microphone (see sec. 3 for a setup description). When the acoustic check is successful, the evaluation layer subsequently initiates a change of the system state to `zoom` mode as well as a zooming of the camera such that a head-shoulder sequence is recorded. When the fixated object does not show acoustic attentive cues or when the object tracking fails, the evaluation layer initiates a state change of one step back from `zoom`- or `focus`-mode to `focus`- or `base`-mode, respectively and initiates the appropriate camera actions. A switch between previous states allows the system to deal with a change of the speaker as well as with situations typical for dynamic scenes such as occlusions or changes in lighting conditions. A similar technique has been proposed by [9]. As pointed out there, the technique of switching between consecutive states adds a considerable amount of robustness to tracking systems in general.

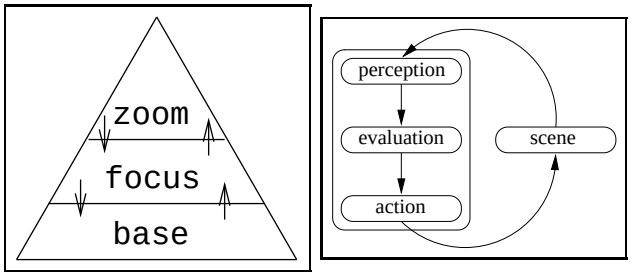

Figure 2: The left image shows the "gazing pyramid": The very bottom represents the `base` state, that allows an overview of the scene. The very top of the pyramid represents the `zoom` state, which allows to record a head-shoulder view of a speaker. The evaluation layer is allowed to switch between successive states. The right graph summarizes the dependencies of the different layers *perception*, *evaluation*, *action*. The action affects the scene as it is perceived by the ACS and thus affects the perception.

2.1 Perception Layer

The camera image of the scene is evaluated first by the *perception layer* in order to detect attentive cues.

When the system is in `base` mode, only visual attentive cues are detected. Visual attentive cues are defined here by skin tone colour and motion within the scene. An *attention image* A indicates where visual attentive cues were found. Colour cues within the scene are represented by colour blobs. Each colour blob i is defined by its mean $\mathbf{p}_i$ and its covariance matrix $\mathbf{C}_i$ and is interpreted as a single object of interest.

In `focus` mode, colour cues and acoustic cues are detected. Acoustic cues result in a binary acoustic detector T that indicates the presence or absence of voices. Colour blobs do not need to be calculated in `focus` mode as this has already been done in the `base` mode. Colour is used for tracking a specific object that is represented by its colour blob. Mean and covariance of the blob is updated as needed.

In `zoom` mode, acoustic cues are detected and blob parameters for the corresponding colour cues are updated as needed. The underlying assumption of the

Figure 3: The left image shows a face that is tracked during `zoom` mode. Right, the corresponding binary attention image A is depicted.

system design is that a speaker emphasizes his/her arguments using more extensive movements while listeners primarily move more slowly and show much less movement. Therefore, the visual attentive cues are defined as *large and fast* movements while *small and slow* movements are not perceived as attentive motion cues. The differentiation between *large and fast* movements and *small and slow* movements is established in two steps: a *selection step* and a *verification step*.

In the selection step, the motion cues are detected by applying special filtering techniques. Fast and large movements are computed by first subsampling and smoothing the camera images I_t at time t in position while keeping the temporal sampling rate at a constant speed of 20 fps (for more details we refer the reader to [1]). Then, the inter-frame difference, $D_t = I_t - I_{t-1}$, is calculated. Subsampling in position while keeping temporal sampling rate high results in a difference image D_t that is insensitive to slow and small movements but sensitive to fast and large movements.

In the second step, the verification step, the difference image D_t is statistically verified using a maximum likelihood estimator [11]. The ML-estimator favours large regions of visual attentive cues in the attention image while very small regions, that tend to be erroneous, are eliminated.

The application of an ML-estimator has two advantages: First, the fast deterministic motion results in a large, approximately coherent area in the binary image A. Secondly, no thresholds are needed which avoids adaptation of the system to different situations.

An example of an attention image is depicted in figure 3: The left image shows a face that is attended by the ACS, the right image shows the corresponding binary attention image A.

For the perception of colour cues, the video images are converted to the HSV colour system. At system start the user is prompted to position a rectangle around his/her inner face. The facial HSV-colour information is then stored in a 1D-histogram, pixels with a value larger than a maximal threshold or lower than a minimal threshold are discarded.

During operation, the histogram is used to calculate for each pixel the probability of showing skin colour. Colour blobs are calculated by first positioning a small window w of some size s over the region of interest. Within this window w and on the basis of the skin colour probabilities we repeatedly calculate the zeroth and the first moments. With respect to these two values the window w is repositioned and resized. This process is repeated until convergence is reached i.e. the position and size of window w does not change any more.

An example for the detection of colour cues is shown in figure 4. The left image shows a person while the ACS is in `zoom` mode. The right image shows the colour cues that are used by the evaluation layer to compute the mean and the covariance.

Voices are understood here as acoustic attentive cues which are perceived by the directional microphone. These cues are only detectable and evaluated if the visual axis of the microphone is already directed towards the noise source so that the microphone perceives acoustic cues only when that person is talking. Acoustic cues are integrated over a time range of about two seconds. The result is stored in an acoustic detector T.

2.2 Evaluation Layer

The resulting binary information of A and T and the colour blob parameters $\mathbf{p}_i$ and $\mathbf{C}_i$ are passed to the *evaluation layer*. The result of the computation of the evaluation layer is a 3D orientation vector $\mathbf{o}$, that is passed to the action layer and which is needed to guide the camera (pan, tilt, and zoom values).

The processing of the evaluation layer depends on the present mode of the ACS:

- In `base` mode only visual attentive cues are evaluated. The attention image A and the colour blobs are evaluated with a ML-estimator as given by

$$s_i = \sum_{\mathbf{x}} A_t(\mathbf{x}) G(\mathbf{x} - \mathbf{p}_i, \mathbf{C}_i), \qquad (1)$$

  for each blob i where G is a Gaussian. The ML-estimations s_i are sorted in decreasing order. The orientation $\mathbf{o}$ is given with respect to the mean and the covariance of the person i that attracts the most attention. The first two components of vector $\mathbf{o}$ define pan and tilt of the camera. The last component of $\mathbf{o}$ defines the zoom position, which, however, is left unchanged in `base` mode.

  Depending on the cues detected, the evaluation layer either keeps the system in `base` mode or initiates a change to `focus` mode.

- In `focus` mode the evaluation layer evaluates the acoustic detector T. During the perception of

acoustic cues the evaluation layer ensures a tracking of the selected candidate by keeping the mean of the corresponding colour blob approximately in the focus of attention. For this, the appropriate camera orientations $\mathbf{o}$ are continuously calculated and passed to the action layer.

When no acoustic cues are detectable, the evaluation layer initiates a change back to `base` mode in order to select another candidate to be attended.

When the system is in `focus` mode and has already detected successfully related acoustic cues, the evaluation layer changes the system's state to the `zoom` mode. In this case the corresponding covariance matrix $\mathbf{C}_{s_i}$ of the corresponding colour blob is used to determine the zoom position of the camera such that the zoomed person covers about $20 - 30\%$ of the camera image.

To summarize: When visual attentive cues are detected in `base` mode, each of the sorted candidates s_i are tested for acoustic cues by switching between `base` mode and `focus` mode until a particular person can be clearly identified as showing considerable visual as well as acoustic cues.

- In `zoom` mode the computation is decomposed in two steps. In the first step a camera orientation $\mathbf{o}$ is calculated from the blob parameters in order to establish tracking.

  In the second step it is decided whether the persons has stopped talking by evaluating the acoustic detector T. In the absence of acoustic cues the evaluation layer initiates a change to the `focus` mode in order to redetect the person that was possibly lost during tracking. If (s)he has stopped talking, the system finally returns to the `base` mode.

In `base` mode, it may happen that an object other than a person results in a colour blob. However, the ACS will not focus this object as it shows neither considerable movement nor significant vocal sound. Therefore, the attention image A will be zero in this region so that the sum within the brackets of eq. 1 will also be zero.

2.3 Action Layer

The *action layer* executes the different system actions as instantiated by the evaluation layer. It controls the orientation as well as the zoom of the camera.

To successfully control the camera for tracking, an internal camera model $\mathbf{m}$ is used. This model is represented by a 3D vector where each component represents one degree of freedom of the camera. The camera model is used to represent the current state

Figure 4: The left image shows a zoomed person while the right image shows the detected skin toned colour.

of the camera. The orientation vector $\mathbf{o}$ is filtered using a Kalman filter $\mathcal{K}$ [5, 12]. Subsequently the filtered orientation vector $\mathcal{K}(\mathbf{o})$ is compared with the camera model. Each degree of freedom of the camera is updated if the difference between the components of the camera model $\mathbf{m}$ and the corresponding components of the filtered orientation vector $\mathcal{K}(\mathbf{x})$ is higher than a certain threshold. The computation speed of 25 fps of the ACS assures that the focussed person is not lost under common video conference conditions. It might happen, that the camera loses the speaker. However, this situation can be easily detected by the system. In this case, the system state is switched back to the `focus` mode, the camera zooms out, and the system is able to redetect the speaker from the wide angle view.

3 Realization and Experiments

In this section we delve into some implementation details about our ACS. A sample setup can be seen in fig. 5 (top left). The ACS requires neither additional hardware nor do users need any special equipment such as tape marks. A Sony EVI-D31 RGB camera with two mechanical controllable degrees of freedom and a zoom range of 5.4 mm – 64.8 mm is used. The camera is connected to a 167 MHz Sun Ultra Sparc via a serial port with a S-bus Sun Video frame grabber card. Using this hardware setup, the system speed reaches 25 fps at a resolution of 384×288 (PAL, half size) pixels per frame. Furthermore, a Beyer MC 737 PV directional microphone is used, which has a sector that covers approximately the width of a face at a distance of 5 meters. The calibration of the acoustic axis of the microphone with the visual axis of the camera has proven to be difficult.

The ACS has been tested with varying illumination conditions and the system has proven to be quite robust to changes in illumination during a conference session. This is due to the robustness supplied by the perception–action cycle [9] and to the normalization of the RGB values [6].

A small demonstration can be seen in fig 5: The top row (from left to right) shows the experimental setup and the wide angle view of the ACS in `base` mode. The middle row shows the motion in

Figure 5: From left to right, top row: Experimental setup, the camera of the ACS is fixed on a tripod in the background; wide angle view of the ACS in `base` mode. Middle row: detected motion in the scene computed by the perception layer; image showing skin toned colours. Bottom row: two sample views of the ACS in `zoom` mode.

the scene that is evaluated by the perception layer (left) and the colour cues (right). The bottom row shows two views of the ACS in `zoom` mode. For an additional demo we refer the reader to our web demo `www.informatik.uni-kiel.de/~vok/vok_{1|2}.mpg`

4 Conclusions

In this contribution an attentive camera system (ACS) for teleconferencing purposes has been presented that is capable to of actively observing a discussion of several persons. The ACS is able to detect speaking persons and to zoom and track them if appropriate attentive cues are visible. The ACS is modeled based on a perception–action cycle (PAC) which allows for modular and flexible construction of a behaviour-based system. Using the PAC approach it is possible to integrate different perception mechanisms such as the detection of colour cues or other attentive cues. This ensures a design of an ACS which is computationally fast and robust.

Acknowledgments

The work has been funded by the Deutsche Forschungs Gesellschaft (DFG), Grants: So 320/1-2 and Ei 322/1-2. R. Herpers acknowledges the support of the DFG, Grant He 2967/1-1. The author would like to thank Joyce P. Wong for her helpful suggestions.

References

[1] E. H. Adelson, C. H. Anderson, J. R. Bergen, J. P. Burt, and J. M. Ogden. Pyramid methods in image processing. *RCA Engineer*, 29(6):33–41, Nov/Dec 1984.

[2] James J. Crowley and Jöelle Coutaz. Vision for man machine interaction. *Robotics and Autonomous Systems*, 19:347–358, 1997.

[3] R. Herpers, M. Michaelis, K.H. Lichtenauer, and G. Sommer. Edge and keypoint detection in facial regions. In *Int. Conf. on Automatic Face- and Gesture-Recognition*, pages 212–217, Killington, Vermont, USA, Oct. 14-16, 1996.

[4] R. Herpers and G. Sommer. An attentive processing strategy for the analysis of facial features. In H. Wechsler et al., editor, *Face Recognition: From Theory to Applications*, pages 457–468. Springer, ASI Series, 1998.

[5] Andrew Kirulata, Moshe Eizenman, and Subbarayan Pasupathy. Predictive head movement tracking using a Kalman filter. *IEEE Trans. Systems, Man, and Cybernetics*, 27(2):326–331, April 1997.

[6] C. H. Lee, J. S. Kim, and K. H. Park. Automatic human face location in a complex background using motion and color information. *Pattern Recognition*, 29(11):1877–1898, 1996.

[7] S. Nayar. Catadioptric omnidirectional camera. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 482–488, Puerto Rico, June 17-19, 1997.

[8] J. Ohya, Y. Kitamura, H. Takemura, F. Ishii, F. Kishino, and N. Terashima. Virtual space teleconferencing: Real–time reproduction of 3d human images. *Journal of Visual Communication and Image Representation*, 6(1):1–25, March 1995.

[9] K. Toyama and G. Hager. Incremental focus of attention for robust visual tracking. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 189–195, 1996.

[10] J.K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sci.*, 13:423–469, 1990.

[11] G. Tziritas and C. Labit. *Motion Analysis for Image Sequence Coding*. Elsevier, Amsterdam, 1994.

[12] Xangdong Xie, Raghavan Sudhakar, and Hanqi Zhuang. Real-time eye feature tracking from a video image sequence using Kalman filter. *IEEE Trans. Systems, Man, and Cybernetics*, 25(2):1568–1577, December 1995.

[13] J. Yang, L. Wu, and A. Waibel. Focus of attention: Towards low bitrate video tele-conferencing. In *Proc. IEEE Int. Conf. on Image Processing*, pages 97–100, Lausanne, Switzerland, Sept. 1996.