And The Fans are Going Wild! SIG plus MIKE

Ian Frank [1], Kumiko Tanaka-Ishii [2], Hiroshi G. Okuno [3,4], Junichi Akita [5]
Yukiko Nakagawa [4], K. Maeda [6], Kazuhiro Nakadai [4], Hiroaki Kitano [4,7]

[1] Complex Games Lab, ETL, Tsukuba, Ibaraki 305-8568, Japan, ianf@etl.go.jp
[2] University of Tokyo, Tokyo 113-8656, Japan, kumiko@ipl.t.u-tokyo.ac.jp
[3] Science University of Tokyo, Chiba 278-8510, Japan, okuno@nue.org
[4] Kitano Symbiotic Systems Project, ERATO, JST, Tokyo 150-0001, Japan, {yuki,nakadai}@symbio.jst.go.jp
[5] Future University, Hakodate, Japan, akita@fun.ac.jp
[6] Chiba University, Chiba 263-8522, Japan, kmaeda@cogsci.l.chiba-u.ac.jp
[7] Sony Computer Science Laboratories, Inc., Tokyo 141-0022, Japan, kitano@csl.sony.co.jp

Abstract. We present an implemented commentary system for real-world robotic soccer. Our system uses an overhead camera to pass game information to the simulator league commentary program MIKE, and uses the humanoid robot SIG to provide a physical embodiment for the commentary. The use of a physical robot allows us to direct audience attention to important events by looking at them, provide more realism to the interaction between MIKE and other humans in the domain, and even set up a dialogue with the audience as part of the commentary itself. Our system combines the multi-modal input of audio and video information to generate a multi-modal commentary that brings the extra dimension of body language to the social interaction that characterises the overall commentary task.

1 Introduction

Good TV or radio commentary of soccer is hard. The action moves fast and the play itself is accompanied by many other events on and around the field. The referee is human and may make mistakes, the players may have rivalries and past histories that result in off-the-ball incidents, the players may argue amongst themselves or with the officials, the attitude of the managers and the team benches may be significant, and the reactions of the crowd can also have an impact on the game.

To date, RoboCup has been successfully used for research on automated commentary by restricting attention to the simulation league [1]. This is still a hard problem, but is much simpler than real-world soccer: information on the player and ball locations is complete and noise-free, there are no linesmen, the referee is just a disembodied decision program, and fouls are limited to free kicks (there are no penalties). In this paper, we describe the first implemented commentary

P. Stone, T. Balch, and G. Kraetzschmar (Eds.): RoboCup 2000, LNAI 2019, pp. 139-148, 2001.

system for the RoboCup robot leagues. This system works by combining two existing research projects: MIKE and SIG.

MIKE is a commentator system developed for the RoboCup's simulator leagues [2]. We adapted MIKE to take as input the log produced by a vision recognition system that processes the video stream from an overhead camera [3]. This 2D representation of the game enables much of the original MIKE code to be retained. In robotic soccer, however, there are many events (such as humans entering the field of play to replace robots) that are too difficult to recognise automatically. We therefore re-designed MIKE as an *assisted* commentary system, which functions as part of a team with one or more humans. MIKE produces commentary by itself as long as it can follow the flow of the game, but when it has difficulty interpreting the scenes, it passes control of the commentary to a human. MIKE is capable of handling a basic dialogue with a human (by simply monitoring voice levels to identify interruptions and determine when it should speak or be quiet), but for our public demonstrations at RoboCup 2000 we used the simpler alternative of supplying MIKE with a repertoire of 40 *episodes* that it could use verbatim to fill in any breaks in the game.

To actually embody MIKE's commentary and situate the computer commentator in the RoboCup environment, we use SIG, a humanoid robot [4]. SIG has binaural microphones and binocular vision and can locate, and direct its gaze at, objects and sounds in 3D space [5]. Using a physical robot to personify the commentary increases the possibilities for realistic commentary. For example, SIG can direct audience attention by looking at the ball or towards the sound of the referee's whistle when it blows, and can add meaning to the commentary (or the speech of the referee or human commentators) by using head movements. It can also monitor, and react to, the level of excitement shown by the crowd watching the game (although this ability was not implemented in time for Melbourne).

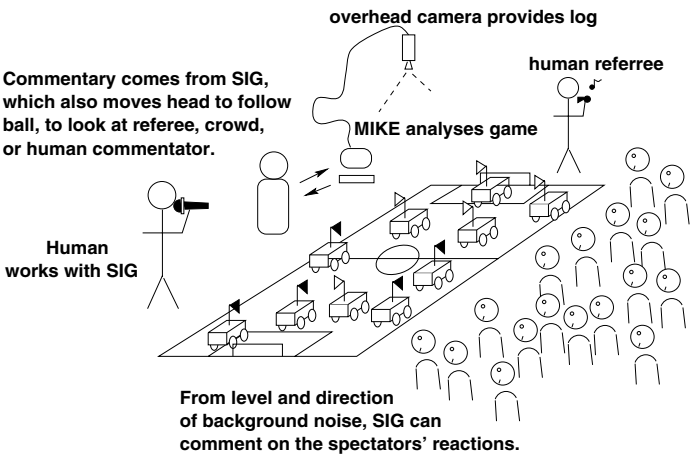

Fig. 1. Pictorial overview of soccer commentary system for real-world robots

Figure 1 shows the overall concept of our commentary system. The rest of this paper describes how we adapted MIKE and SIG towards realising this vision.

2 Overall Design

We work on the principle that the referee, robots, team designers, other commentators, and audience members are all significant actors in the drama of a RoboCup game. To produce an integrated system combining SIG and MIKE with all these other agents, we designed the simple architecture of Figure 2. In this figure, the ball and players' positions are assumed to come from the global vision system initially developed by [3]. Our improved version of this system can capture 30 frames per second on a PentiumIII at 450MHz. We identify robots by finding the five most promising matches for the ping-pong balls of each team, but also carry out error reduction by only looking for robots within the *detection radius* (usually, less than 50cm) of the positions where robots were successfully identified in the previous frame. Our implementation also identifies the extra markers of the robots (without limiting the detection radius) and humans entering the pitch (via an increase in the number of extra markers).

MIKE produces the basic commentary for a game on the basis of the information from this global vision system. Although MIKE was primarily designed to work with the logs of simulated games, a large amount of the basic code was retained (we discuss the changes in §3). The data from the global vision system is also passed to SIG to allow it to easily implement ball and robot tracking. Implementing such tracking using SIG's own cameras is future work.

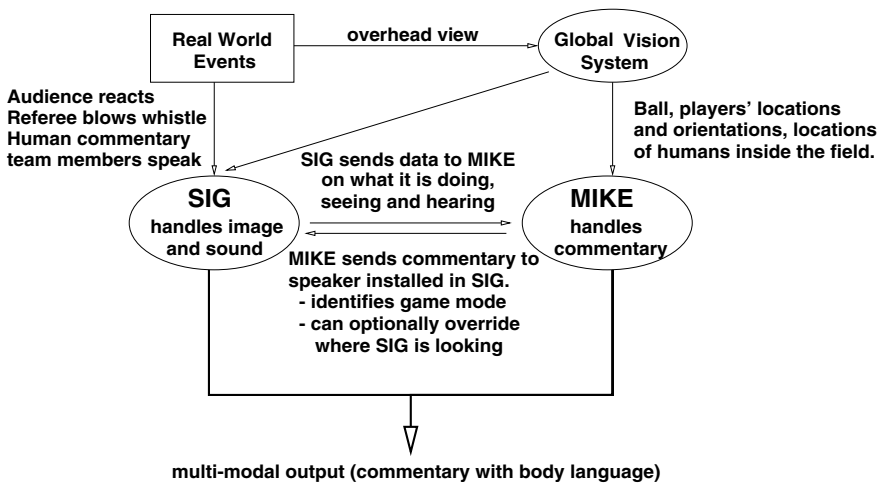

Fig. 2. Architecture combining MIKE and SIG to produce an embodied commentator

	Play mode	Stoppage mode
SIG	• follow ball • look towards sound of whistle • look at speaking commentator • tell MIKE what it does, sees, hears	• look towards sound of whistle • look at speaking commentator • nod and look interested in what human commentator says • look at any people in field of play • tell MIKE what it does, sees, hears
MIKE	• send commentary to speaker • (optional) tell SIG where to look • tell SIG when mode changes to stoppage mode	• Make the decision when to switch mode back to play mode, and inform SIG

Table 1. The roles of SIG and MIKE in play and stoppage modes

To maintain coordination between SIG and MIKE, we separated the roles of the two systems, as summarised in Table 1. At the basic level, SIG is expected to be capable of deciding for itself what it should do as part of a commentary team. However, we define the two distinct modes of *play* and *stoppage*, that affect its behaviour. In play mode, SIG concentrates on following the ball, but may also look at the source of any sounds that it detects (in our current implementation, the referee's whistle). In stoppage mode, however, SIG ignores the ball and looks around at other things in the domain. Since we didn't use human commentators for our demonstrations, SIG could not look towards them, but in future demonstrations we hope to show how SIG can react when human commentators speak, signalling its attention by nodding or tilting its head occasionally.

The responsibility of identifying the correct play mode is given to MIKE. This is actually a complicated task that can be viewed as a first step along the path to automated refereeing. With the ability to recognise whistles (using SIG's microphones) and to identify when humans enter the field of play (from the global vision system), MIKE can identify the game mode reasonably well. During our demonstrations, however, we gave a human operator the option of overriding MIKE's decisions. MIKE also has the task of refining the overall coordination of the commentary by passing SIG explicit directions to focus on specific features of domain. For example, if MIKE decides to make a comment about the level of the crowd noise, it is better for SIG to be looking at the crowd for the duration of this comment.

Overall, the interaction between MIKE, SIG and the other actors in the domain creates an unusual example of human-computer interaction. This interaction is multi-modal in both input and output. First, to perceive the world, data from 3D sound processing is integrated with visual data. Then, the commentary generated from this data is complemented by interaction with humans in the domain. Since we use a physical robot to embody the computer commentary, we

can use body language as a mode of communicating information to the audience. Effectively, each actor in the domain combines with SIG and MIKE in a social interaction that ultimately includes the audience itself as part of the event.

3 Descriptive Content: The nature of the robotic leagues

To apply MIKE — a commentator system for simulated soccer — to a game played by robots, we had to take account of the different characteristics of the real world domain. Table 2 gives an overview of these differences, the most important of which we can summarise below.

- The real robot league is surrounded by a wall.
- The ball tends to move much faster, both because the wall keeps the ball in play and because many teams have developed specialised hardware for kicking the ball hard.
- The higher ball speeds tend to result in the ball bouncing off players more often, and less actual passing and control of the ball
- The robot league has more complicated rules to deal with situations such as robots or the ball becoming stuck. Humans are allowed to enter the pitch, and there can be delays for repairs of robots.
- Robot games can have penalty kicks, with attendant lengthy pauses for the team developers to reposition robots.

When play is actually in progress, the higher ball speed means that it is more effective for MIKE to produce commentary that concentrates more on team formations and on overall indications of which team is attacking. However, to cope with long delays before penalty kicks and for repair of robots, we found that the only reliable way to produce robust commentary was to rely on a collaboration with a human counterpart.

Feature	Simulation league	Small-sized robot league
location data	complete	includes errors
orientation data	complete	includes errors
ball speed	bounded at reasonable level	can move very fast
environment	2D field without wall	3D field surrounded by wall
rules	simple refereeing almost 100% automated no penalty kicks	complicated human referee penalty kicks
time	no stoppages	potentially long stoppages
frames	varies (10 per sec produced by vision system we utilise [3])	30 per sec

Table 2. Differences between simulation and small sized robot leagues

3.1 Assisted Commentary & Episodes

Many of the events surrounding a robot soccer game are too complex to contemplate identifying automatically with today's technology. For example, a referee may show a yellow card to one of the teams, or may decide to explain a decision in English. Since MIKE cannot hope to recognise these events by itself (or even with SIG's assistance), we developed an *assisted* commentary mode for MIKE. In assisted commentary, a human works with MIKE to commentate a game. If MIKE at any time feels that it does not understand what is happening, it can pass control to the human. MIKE then expects the human to explain the situation until the game reaches a state that is easier to understand.

To do this, we extended a recently developed version of MIKE that already works with two commentators. In [6], we describe the general advantages of dividing up the explanation task between multiple agents in real-time domains, and give as an example a version of MIKE that explains soccer games with two separate voices: an *expert* and an *announcer*. We expected the general approach we used for passing commentary between two computer-controlled commentators to also hold when one of the commentators was a human. We therefore developed a simple system that MIKE could use to determine when to speak: monitor a microphone to determine when the human commentator is speaking, then keep silent unless there is a two second pause. Using simple templates such as "Over to you" and "Thanks" to pass the conversation backwards and forwards produced a reasonable effect, but during our public demonstrations we decided to fall back on the more reliable alternative of using MIKE to speak the entire commentary, but allowing the human commentator to assist MIKE at difficult points by selecting *episodes* for it to say.

The episodes that we designed for MIKE were primarily designed to fill the long pauses that occur when teams call time-outs. They can therefore be quite long. However, there are also some short examples that can be used to interrupt the commentary of a game (*e.g.*, "Please, no flash photography", or "The referee awarded a penalty"). There are 41 episodes in total, and a longer example is shown in Figure 3 (commented lines represent commands to SIG, such as pause, nod, or look at a pre-specified location, as described in the following section). A video of SIG speaking the episode of Figure 3 during a break in play at Melbourne is available from the MIKE web pages [7].

4 SIG, the Humanoid

Figure 4 and Figure 5 show the physical appearance of SIG. To give an idea of scale, the height of upper torso is approximately 70cm, and weighs 15Kg. When using SIG to commentate for RoboCup games, it is raised on a platform so that it is slightly lower than the height of an average person. This elevation gives SIG a good view of the game field (on the floor) and also exaggerates the effect of a change of focus from the pitch to other objects such as the referee, crowd, or a human commentator.

```
 This seems to be a long break in the game here,
 so I'm going to take the chance to advertise myself!
#<PAUSE>
 I am making the commentary on this game thanks to the work
 of three research groups. First, the good people at E T L and Tokyo
 University in Japan gave me a voice and the ability to follow a game
 of soccer and understand it. Then, the people at ERATO in Tokyo gave
 me a body that can move...
#<PAUSE>
 Sometimes!
#<PAUSE>
 Finally, we hooked up with some cameras provided by people
 at the Future University in Japan to actually watch these robot games.
#<PAUSE>
 In my first incarnation in 1997, I was called MIKE,
 and I could just commentate on the RoboCup simulation league.
 Now I'm really glad to have this body
#<NOD>
#<PAUSE>
 and to be able to talk to all of you
#<AUDIENCE_L>
 about todays
#<AUDIENCE_R>
 soccer game. I hope you are enjoying my commentary today.
 If not, please don't tell my developers.
```

Fig. 3. Part of an episode used by MIKE

The surface of SIG is covered with urethane. Functionally, this covering reduces noise from SIG's internal mechanical and electrical components, as well as simply protecting them. This noise reduction results in a simplification of the processing for the binaural inputs. Aesthetically, FRP was chosen because it was found to be a good compromise between the functional requirements and the need to appeal (both statically and dynamically) as a believable actor in physical situations.

SIG uses both *active vision* and *active audition* [5] in the sense that it moves its body or changes camera parameters to perceive objects or sound sources better. SIG has a total of four degrees of freedom in moving its head and neck, which it manipulates using four DC motors with potentiometers. In turn, each eye has two degrees of freedom (pan and tilt), and also a focus and zoom control. For audio information, four microphones are configured as a couple of binaural microphones and a set of internal microphones, located just inside of the ear recesses. The visual and auditory processing abilities of SIG can be summarised as follows [8]:

1. The ambiguity in the sound source direction obtained by vision and audition is $\pm 1^\circ$ and $\pm 10^\circ$, respectively. Thus, visual information on a sound source is

Fig. 4. The humanoid, SIG

Fig. 5. Close-up of SIG's head

used in preference to auditory information when possible. Note that a typical human's auditory capability is about $\pm 8°$.
2. When an exact direction of sound source is available, a direction-pass filter separates sounds originating from the specified direction.
3. When a sound source is out of sight or occluded, its auditory direction is used for improving visual tracking.

SIG was designed to tackle one of the main challenges of active audition: the separation of target sounds from the overall mixture of sounds reaching the robot. However, in our soccer domain, such 'sound source separation' is not the primary goal. Rather we are happy simply with the ability to identify auditory events such as the referee's whistle, and audience applause. For the demonstration in Melbourne, we implemented the following abilities:

1. Move two eyes and/or its body to track the ball or some robot player based on information from the overhead camera (this was easier than using SIG's own visual tracking system) or based on commands from MIKE specifying what SIG should do to best convey the current commentary.
2. Move two eyes and/or its body to face pre-programmed locations, such as the left or right side of the audience (see examples in Figure 3).
3. Identify when the whistle is blown (using a template of the whistle sound).
4. Nod and tilt its head based on commentary content.

Although these abilities ignore the input of the audience and human commentators (we did not use human commentators in our demonstrations) they were still enough to allow us to substantially increase the level of audience involvement in a game. For example, one of the episodes used by MIKE was "Are there any $\langle Team \rangle$ fans in the audience?" After directing SIG to look at the audience for a short while, MIKE would then randomly pick an assessment of the

audience reaction (*e.g*, "Good", "I see", "OK", "Much Better"). This is a very simple level of interaction, but the audience enjoyed these exchanges, even when (or, especially when) MIKE's comments did not match the level of cheering in the audience response. This was a simple but convincing demonstration of the potential of the "SIG plus MIKE" combination to become a genuine participant in a social interaction involving a very large number of agents.

5 Related work

In terms of soccer commentary itself, we mentioned in the Introduction that all the efforts within RoboCup itself have concentrated on the simulation league. However, SOCCER, the forerunner of one of these RoboCup systems was actually designed to interpret short sections of video recordings of real soccer games [9]. To tackle the machine vision and scene interpretation problems involved in this task, SOCCER used *geometrical scene descriptions* [10] to understand the images. SOCCER also employed multimedia presentation techniques to combine text, graphics and video in its presentations of these recordings.

Our work differs from SOCCER in that our commentator has to handle *entire* matches from start to finish, rather than just short sections. This means there are many more problems of discourse continuity, management of repetition, and the handling of stoppage time within a game. Our commentator is also physically embodied and actually situated in a real soccer game, thus also increasing the number and type of domain events to comment on, and necessitating the use of both audio information and visual information to understand a game.

One novel consequence of our use of an embodied commentator is the potential it offers for social interaction with the other actors in the domain, such as the audience members, other commentators, and referee. Brooks [11], for instance, has pointed out that "robots with humanoid form are a tool for investigating and validating cognitive theories". The SIG plus MIKE model can allow us to investigate how social interactions can be managed — and how they are perceived — in a system with a large number of agents. For instance, in terms of the actual interaction between SIG and human commentators, we could hope to extend the results reported by the developers of Waseda University's robot that controlling gaze is a way to produce "confirmation of communication" [12]. The use of gaze can help not only during dialogues between SIG and other humans, but can also be used (as well as other forms of body language) to actively pass information to the audience. We especially hope to implement an 'interview' mode for our system, where SIG is placed *between* two human speakers, so that it can turn its head to face each one as they take turns to talk. This kind of interaction would be ideal for increasing the impact when a human commentator interviews one of the 'personalities' connected with a game, such as a developer of one of the competing teams.

6 Conclusions

We have shown how SIG and MIKE can be combined to process multiple modes of information and unify some of the many threads of a robotic soccer game into a single, multi-modal, interactive commentary. We demonstrated our implementation of this system at four separate games of the 2000 RoboCup in Melbourne, where we received favourable feedback (and made the audience laugh). There is still much work to be done to produce a genuine autonomous commentator for this domain, but our work represents a significant first step, and proves the feasibility and the value of the challenge.

References

1. E. Andrè, K. Binsted, K. Tanaka-Ishii, S. Luke, G. Herzog, and T. Rist. Three RoboCup simulation league commentator systems. *AI Magazine*, 21(1):57–66, Spring 2000.
2. K. Tanaka-Ishii, I. Noda, I. Frank, H. Nakashima, K. Hasida, and H. Matsubara. MIKE: An automatic commentary system for soccer. In *Proceedings of ICMAS-98*, pages 285–292, 1998.
3. J. Akita, J. Sese, T. Saka, M. Aono, T. Kawarabayashi, and J. Nishino. Simple soccer robots with high-speed vision system based on color detection hardware. In *RoboCup-99 Workshop, Stockholm, Sweden*, pages 53–57, 1999.
4. H. Kitano, H. G. Okuno, K. Nakadai, I. Fermin, T. Sabish, Y. Nakagawa, and T. Matsui. Designing a humanoid head for robocup challenge. In *Proceedings of Agent 2000 (Agent 2000)*, page *to appear*, 2000.
5. K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano. Active audition for humanoid. In *Proceedings of AAAI-2000*, page *to appear*, Austin, TX, 2000.
6. K. Tanaka-Ishii and I. Frank. Multi-agent explanation strategies in real-time domains. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 158–165, Hong Kong, 2000.
7. MIKE web page. `http://www.etl.go.jp/etl/suiron/~ianf/Mike`.
8. H.G. Okuno, Y. Nakagawa, and H. Kitano. Integrating auditory and visual perception for robotic soccer players. In *Proceedings of International Conference on Systems, Man, and Cybernetics (SMC-99)*, volume VI, pages 744–749, Tokyo, Oct. 1999. IEEE.
9. E. Andrè, G. Herzog, and T. Rist. On the simultaneous interpretation of real world image sequences and their natural language description: The system soccer. In *Proc. of the 8th ECAI*, pages 449–454, Munich, 1988.
10. B. Neumann. Natural language description of time-varying scenes. In D.L. Waltz, editor, *Semantic Structures: Advances in Natural Language Processing*, pages 167–207. Lawrence Erlbaum, 1989. ISBN 0-89859-817-6.
11. R. A. Brooks, C. Breazeal, R. Irie, C. C. Kemp, M. Marjanović, B. Scassellati, and M. M. Williamson. Alternative essences of intelligence. In *Proceedings of AAAI-98*, pages 961–968, Madison, WI, 1998.
12. H. Kikuchi, M. Yokoyama, K. Hoashi, Y. Hidaki, T. Kobayashi, and K. Shirai. Controlling gaze of humanoid in communication with human. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 1, pages 255–260, Victoria, Canada, 1998. IEEE.