Automatic Camera Control System for a Distant Lecture Based on Estimation of Teacher’ s Behavior

Shimada, Atsushi
Department of Intelligent Systems, Kyushu University

Suganuma, Akira
Department of Intelligent Systems, Kyushu University

Taniguchi, Rin-ichiro
Department of Intelligent Systems, Kyushu University

http://hdl.handle.net/2324/4031

出版情報：Proceedings of the Seventh IASTED International Conference on Computers and Advanced Technology in Education, pp.106-111, 2004-08
バージョン：
権利関係：

**Proceedings of the Seventh IASTED International Conference
Computers and Advanced Technology in Education
August 16-18, 2004, Kauai, Hawaii, USA**

AUTOMATIC CAMERA CONTROL SYSTEM FOR A DISTANT LECTURE BASED ON ESTIMATION OF TEACHER'S BEHAVIOR

Atsushi Shimada
Department of Intelligent Systems
Kyushu University
6–1 Kasuga-koen, Kasuga, 816–8580, Japan
email: atsushi@limu.is.kyushu-u.ac.jp

Akira Suganuma
Department of Intelligent Systems
Kyushu University
6–1 Kasuga-koen, Kasuga, 816–8580, Japan
email: suga@limu.is.kyushu-u.ac.jp

Rin-ichiro Taniguchi
Department of Intelligent Systems
Kyushu University
6–1 Kasuga-koen, Kasuga, 816–8580, Japan
email: rin@limu.is.kyushu-u.ac.jp

ABSTRACT
We are developing an Automatic Camera control system for Education: ACE, which captures a lecture using both a blackboard and a screen. ACE focuses on an oblect explained by a teacher. When this recording strategy is realized, it is necessary for ACE to extract a teacher's behavior and his/her explaining object. In this paper, we describe our algorithm to estimate a teacher's behavior by image processing and the camera control strategy to take suitable shots. We have applied ACE to recording a real lecture to validate it.

KEY WORDS
Distant lecture, Image processing, Estimation of teacher's behavior, Extraction of explained area, Camera control strategy

1 Introduction

The growth of a communication network technology enables people to take part in a distant lecture. When lecture scenes for the distant lecture are captured, a camera-person usually controls a camera to take suitable shots; alternatively, the camera is fixed and captures the same location all the time. It is not easy, however, to employ a camera-person for every occasion, and the scenes captured by a fixed camera hardly gives us a feeling of the live lecture. It is necessary to control a camera automatically.

We are developing a supporting system for a distant lecture, which estimates teacher's behavior and controls an active camera to take a suitable shot. We call it "ACE" (Automatic Camera control system for Education). ACE captures an area which a teacher is explaining. The previous version of ACE[1, 2] supports a traditional-style lecture in which a teacher uses only blackboard. Since a teacher frequently explains objects written on the blackboard in a traditional-style lecture, the previous version of ACE could take suitable shots by focusing on the latest object written on the blackboard. Nowadays, a teacher teaches his students with some visual facilities. Many new-style lectures which the teacher uses both a blackboard and a video projector are held in many universities. The previous version of ACE cannot take suitable shots for the lecture in such a style because it is not always true that the area which a teacher is explaining is the latest object written on the blackboard.

When a camera-person controls a camera, he/she takes an important scene in a lecture. Both objects written by a teacher on a blackboard and the area that he/she is explaining are important. A camera-person appropriately changes an area to capture (target area) according to a lecture scene. We guess he/she changes the target area when a teacher's behavior changes. For example, when a teacher is writing an object on the blackboard, a camera-person captures with a focus on the teacher and the object. When a teacher is explaining to his/her student's, the camera-person captures the explained area. ACE needs to estimate a teacher's behavior for the purpose of capturing the similar shot which a camera-person took. In this paper, we describe our algorithm to estimate a teacher's behavior by image processing and the camera control strategy to take a suitable shot.

2 Overview of ACE

2.1 Design

A style of the distant lecture which we envisage is illustrated in Figure 1. A teacher teaches his/her students in a local classroom, and students in remote classrooms take part in the lecture by watching the video captured in the local classroom. ACE supports the lecture in which a teacher teaches his/her students by using both a blackboard and a screen.

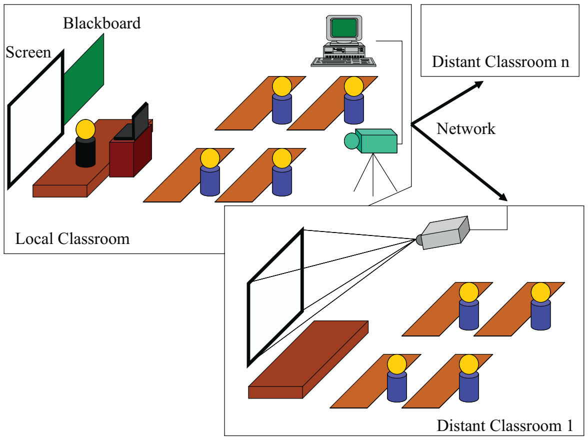

Figure 1. Lecture style assumed in ACE

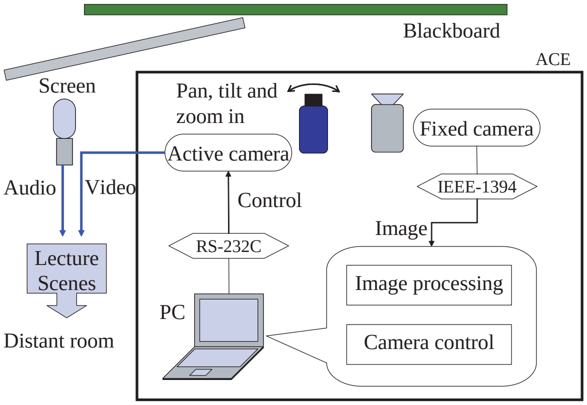

Figure 2. Framework of ACE

2.2 Framework of ACE

Figure 2 shows the framework of ACE. ACE needs two cameras. One is a fixed camera, which captures whole platform for image processing. The other is an active camera to capture a suitable shot, and its video is transmitted to remote classrooms. The image captured by the fixed camera is sent to PC (in Figure 2) via an IEEE-1394. The PC analyzes the image and controls the active camera via an RS-232C. The video captured by the active camera and the audio picked up by a microphone are sent to remote classrooms via the network using DVTS (Digital Video Transport System)[3] software.

2.3 Camera Control Strategy

What does ACE capture? One solution for this problem is to take the scene that students want to watch, but many scenes are probably requested by many students at the same time. Although this solution needs the consensus of all students, it is very difficult to make it. We have decided, therefore, that ACE captures the most important thing from a teacher's point of view. When we designed the previous version of ACE, we assumed that the most important thing is the latest object written on the blackboard. The previous version of ACE took a shot zoomed in on the object after the teacher had written it on the blackboard. After a-few-second zooming, the previous version of ACE zoomed out and took a shot containing the latest object and a region near it. However, this strategy has some problems. The previous version of ACE cannot take a suitable shot when the teacher explains object written before. The lecture scene captured by the previous version of ACE changes at short intervals because the latest object is often found when he/she goes on writing objects on the blackboard for a long periods of time. Such a video is not appropriate for students who take part in the distant lecture. Particularly the assumption in the previous version of ACE is not appropriate for new-style lectures because the teacher don't always explains the latest object.

We have adopted, therefore, the strategy that ACE captures an object explained by a teacher. When the teacher is writing object on a blackboard, ACE captures the teacher and the object. When he/she is explaining to his/her students, ACE captures the latest object. If he/she is explaining the objects written on the blackboard before, ACE also captures them. On the other hand, the image processing component on the PC checks whether the next target which ACE should capture is included in the current capturing area or not. If the next target is included, ACE need not move the active camera. This solves the problem that the lecture scene changes at short intervals. When he/she is explaining the objects on the screen, ACE captures whole of the screen even if the objects are anywhere on the screen.

3 Modeling of Teacher's Behavior

3.1 Teacher's Behavior in a Lecture

We observed some lecture videos to investigate the teacher's behavior. . We found out that the behavior of the teacher could be categorized into three kinds: "Writing", "Explaining" and "Moving". When the teacher was writing some objects on the blackboard, we categorized his/her behavior as "Writing". When the teacher was explaining objects on the blackboard or on the screen, we categorized his/her behavior as "Explaining". We categorized the other kind of behavior as "Moving".

3.2 Creating Teacher's Behavior Model

We got the position of the teacher's centroid $\boldsymbol{g}(t) = (g_x(t), g_y(t))$, face $\boldsymbol{f}(t) = (f_x(t), f_y(t))$ and hand $\boldsymbol{h}(t) = (h_x(t), h_y(t))$ from the one of the lecture videos by the hand work in 2 fps. The positions may be represented as a time series $\boldsymbol{I}(0), \boldsymbol{I}(1), \cdots, \boldsymbol{I}(T)$, where $\boldsymbol{I}(t)$ denotes the position of the teacher's centroid, face, and hand at time t.

$$\boldsymbol{I}(t) = (g_x(t), g_y(t), f_x(t), f_y(t), h_x(t), h_y(t)) \quad (1)$$

We made 5-dimensional feature vector $\boldsymbol{x}(t) = (v_1(t), v_2(t), v_3(t), v_4(t), v_5(t))^T$ by using $\boldsymbol{I}(t)$ and $\boldsymbol{I}(t-1)$. The feature vector has following five elements.

$$v_1(t) = |f_x(t) - f_x(t-1)| \quad (2)$$

$$v_2(t) = |f_y(t) - h_y(t)| \quad (3)$$

$$v_3(t) = \sqrt{(f_x(t) - h_x(t))^2 + (f_y(t) - h_y(t))^2} \quad (4)$$

$$v_4(t) = |g_x(t) - h_x(t)| \quad (5)$$

$$v_5(t) = |g_y(t) - h_y(t)| \quad (6)$$

ACE has to categorize an unknown feature vector. We have decided to use the stochastic approach. We can regard the feature vectors as the distribution of the points over the vector space. We use the Gaussian mixture model in order to approximate the distribution. The Gaussian mixture modeling approximates a probability density function by a weighted sum of multivariate Gaussian densities[4]. Gaussian mixture model with n components can be used to approximate the joint density function $p(\boldsymbol{x})$ using

$$p(\boldsymbol{x}) = \sum_{i=1}^{n} w_i g(\boldsymbol{x}; \mu_i, \Sigma_i) \quad (7)$$

$g(\boldsymbol{x}; \mu, \Sigma)$ is the multivariate Gaussian density function over the space of $\boldsymbol{x}$, and in this instance, $\boldsymbol{x} = (v_1, \cdots, v_5)^T, d = 5$. The multivariate Gaussian density function is defined by

$$g(\boldsymbol{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \mu)^T \Sigma^{-1} (\boldsymbol{x} - \mu)\right\}, \quad \boldsymbol{x} \in \boldsymbol{R}^d \quad (8)$$

with mean vector μ and covariance matrix Σ, and thus the entire mixture is defined by the parameter set $\theta = \{w_i, \mu_i, \Sigma_i | i = 1, \cdots, n\}$.

Since it is not possible to directly obtain optimal parameter values for a Gaussian mixture model, it is necessary to use some form of numerical optimization scheme. We applied, therefore, EM-algorithm to estimate the parameter set of the Gaussian mixture model. We got three model parameter sets (θ_W, θ_E, θ_M). The parameter set θ_W was estimated from 1,463 data, θ_E was estimated from 1,248 data, and θ_M was estimated from 857 data.

4 Image Processing

ACE analyzes the image captured by the fixed camera to acquire following four information.

Teacher's region
The region which encloses the teacher.

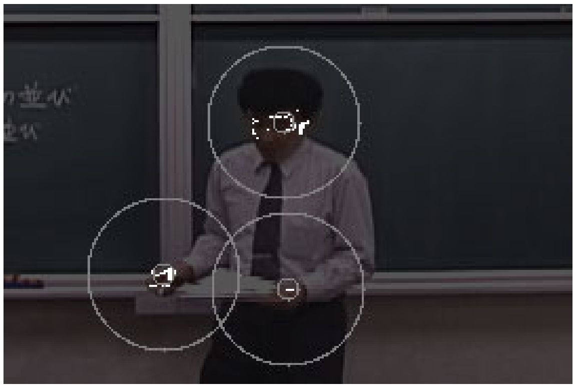

Figure 3. The result of clustering

Teacher's behavior
The behavior which the teacher is doing; "Writing", "Explaining", or "Moving".

Object region
The region which encloses the object written by the teacher on the blackboard.

Explained area
The area which the teacher is explaining.

4.1 Detecting Teacher's Region

ACE has to segment a teacher. We use the background subtraction technique to detect objects in the image. The background image is captured before opening the lecture. The image contains the screen and the blackboard on which no object was written. After subtracting the background from the image captured by the fixed camera during the lecture, ACE can get a foreground image, which consists of objects written on the blackboard, the teacher and so on. We would like to detect only the teacher. We apply, therefore, the erosion technique to the foreground image. We use a 5×5 mask because the objects or the noise in the foreground image are thinner and smaller than the teacher's body. After the erosion, our system makes the histogram of all highlight pixels in the filtered image because some noises are still remained in the image. ACE extracts the teacher's region from the histogram by setting an appropriate threshold.

4.2 Estimating Teacher's Behavior

After detection of the teacher's region, ACE extracts feature points $\boldsymbol{I}$(t). ACE acquires the centroid of the teacher of the foreground image as the center of the highlight pixels in the teacher's region. The teacher's face and hands are detected by extracting skin color pixels in the teacher's region. We used HSV color space to extract skin color pixels. The teacher's hand is often hidden by his/her own body, so our system sometimes detects his/her both hands and sometimes detects only one hand or no hand. Our system

Table 1. The parameter for the camera control

- Teacher's behavior
- Teacher's region
- Explained area
- The time when ACE controled the zoom rate of the active camera at last
- The area which ACE is capturing

categorizes, therefore, the skin color area into at most three clusters, because the skin color area consists of three parts, by applying the k-means clustering method. Figure 3 is the result of clustering. The white colored circles show the clusters. ACE needs to categorize the clusters into his/her face and hands. We use the Hough transform to detect a face because the shape of the face can be approximated as a circle. The face position, which was detected by the Hough transform, is compared with each cluster. ACE categorizes the nearest cluster as a facial area. The other clusters are regarded as teacher's hands. ACE calculate the feature vector by using the feature points which were gotten above. ACE puts the feature vector into the Gaussian mixtures which indicate each behavior model, and acquire probability of each behavior. The behavior whose probability is highest is regarded as the teacher's behavior at time t.

4.3 Extracting the Latest Object

ACE detects the object on the blackboard as the foreground image by the background subtraction technique. However, the teacher is also contained in the foreground image. ACE extracts the latest object on the blackboard masking out the teacher's region.

ACE records following two information when it extracts the latest object.

- The position where the latest object was written.
- The time when the latest object was extracted.

4.4 Extracting Explained Area

ACE extracts an explained area by using three information; the teacher's region, the teacher's behavior and the object region.

Explained area 1

When the teacher's behavior is estimated as "Writing", ACE refers to the time when the object region was written and extracts all of the object regions which had been written in the past since t_W seconds as the explained area.

Explained area 2

When the teacher's behavior is estimated as "Explaining", ACE extracts all of the object regions which had been written in the past t_E seconds as the explained area. By setting $t_W < t_E$, ACE extracts the explained area more widely than that extracted when the teacher is "Writing" .

Explained area 3

When the teacher's behavior is estimated as "Explaining" and he/she is pointing an object written on the blackboard, ACE extracts all of the object regions which are on the vector from his/her centroid to hand as the explained area.

Explained area 4

When the teacher's behavior is estimated as "Explaining" and he/she is near the screen, ACE extracts whole of the screen as the explained area.

4.5 Deciding Target Area

ACE finds a target area according to a parameter described in Table 1.

ACE changes the zoom rate of the active camera according to the teacher's behavior. If the interval between the times when ACE changed the zoom rate is shorter than a threshold, it doesn't change the zoom rate. In the case of changing the zoom rate, ACE records the time. If the next target area is "Explained area 4", ACE changes the zoom rate unconditionally.

According to above condition, when ACE changes the zoom rate, it captures the target area at the new zoom rate. On the other hand, when ACE don't change the zoom rate, it checks that the next target area is included in the area which it is capturing now. If the area is not included, ACE controls the active camera with only pan-tilt to capture it.

5 Experiment

5.1 Applying ACE to a real lecture

We have developed ACE and done an experiment of applying ACE to a real lecture. We delivered two 20-minutes lectures for 58 undergraduates. A teacher taught them by using both a blackboard and a screen. We took the lecture scene with ACE and made two camera-persons who can understand the lecture take a scene of the same lecture. In our experiment, these shots were not transmitted over the network but were recorded on video cassettes and played in the classroom with VCR.

After each video lecture, we had the students fill in a questionnaire which consists of following eight questions;

(1) Could you watch the teacher's action well?

(2) Could you watch the objects on the blackboard well?

(3) Could you watch the object you wanted?

Table 2. Results of the questionnaire

No.	Score	Person	ACE
(1)	Average	2.90	2.93
	5: Excellent	0.00%	0.00%
	4: Good	32.76%	34.48%
	3: Satisfactory	31.03%	31.03%
	2: Unsatisfactory	29.31%	27.59%
	1: Poor	6.90%	6.90%
(2)	Average	2.19	2.09
	5: Excellent	0.00%	0.00%
	4: Good	8.62%	5.17%
	3: Satisfactory	34.48%	22.41%
	2: Unsatisfactory	24.14%	48.28%
	1: Poor	32.76%	24.14%
(3)	Average	2.66	2.41
	5: Excellent	1.72%	0.00%
	4: Good	18.97%	17.24%
	3: Satisfactory	37.93%	25.86%
	2: Unsatisfactory	25.86%	37.93%
	1: Poor	15.52%	18.97%
(4)	Average	2.71	2.34
	5: Excellent	0.00%	0.00%
	4: Good	27.59%	17.24%
	3: Satisfactory	27.59%	20.69%
	2: Unsatisfactory	32.76%	41.38%
	1: Poor	12.07%	20.69%
(5)	Average	2.22	1.97
	5: Excellent	0.00%	0.00%
	4: Good	15.52%	0.00%
	3: Satisfactory	32.76%	13.79%
	2: Unsatisfactory	43.10%	68.97%
	1: Poor	8.62%	17.24%
(6)	Average	2.55	2.17
	5: Excellent	0.00%	0.00%
	4: Good	5.17%	0.00%
	3: Satisfactory	27.59%	17.24%
	2: Unsatisfactory	51.72%	62.07%
	1: Poor	15.52%	13.79%
(7)	Average	2.71	2.62
	5: Excellent	3.45%	0.00%
	4: Good	25.86%	24.14%
	3: Satisfactory	22.41%	31.03%
	2: Unsatisfactory	34.48%	27.59%
	1: Poor	13.79%	17.24%
(8)	Average	3.28	2.53
	5: Excellent	3.45%	0.00%
	4: Good	39.66%	10.34%
	3: Satisfactory	41.38%	37.93%
	2: Unsatisfactory	12.07%	46.55%
	1: Poor	3.45%	5.17%

Table 3. Result of our comparing

	Person 1	Person 2
Similar	82.08%	89.33%
Not similar	17.92%	10.67%
Case (a)	3.75%	0.58%
Case (b)	2.50%	3.50%
Case (c)	3.08%	2.92%
Case (d)	3.92%	0.67%
Case (e)	4.67%	3.00%

Table 4. Result of evaluating

	Former 20-minutes		Later 20-minutes	
	Person 1	ACE	Person 2	ACE
Suitable	93.33%	82.42%	91.00%	84.50%
Unsuitable	6.67%	17.58%	9.00%	15.50%
Case (A)	0.00%	1.08%	0.00%	0.50%
Case (B)	5.67%	0.75%	6.17%	9.58%
Case (C)	0.67%	5.75%	2.83%	0.92%
Case (D)	0.33%	1.83%	0.00%	0.92%
Case (E)	0.00%	8.17%	0.00%	3.58%

(4) Were you given a feeling of the live lecture?

(5) Could you give the scene a overall score as a lecture one?

(6) Did you understand the content of this video lecture rather than that of the normal lecture?

(7) Could you watch the objects on the screen well?

(8) How about the camera motion?

They scored each scene from 1 to 5. The distribution of the scores of each question is shown in Table 2. In this table, the scores in “Person” column are ones of the scene captured by the camera-person, and the scores in “ACE” column are ones of the scene captured by ACE.

We applied the t-test to compare the scores of the camera-person and ACE. Our null hypothesis is “*The scene captured by ACE is as good as the scene captured by the camera-person.*” This null hypothesis is rejected with 5% level of significance in questions (5) and (8). In questions (1), (2), (3), (4), (6) and (7), the lecture scene captured by ACE is as good as the scene captured by the camera-person. In the questions (5) and (8), however, the evaluation of ACE is inferior to that of the camera-person. The camera-person can move the camera smoothly. ACE cannot move the camera as well as the camera-person because of the pan-tilt unit restriction.

5.2 Comparing Two Shots

We investigated two video lectures; one is the scene captured by the camera-person, the other is the scene captured by ACE. We judged whether the scene captured by ACE had been similar to the one captured by the camera-person. We regarded following cases as different.

- The object captured by ACE is different from the one captured by the camera-person.
- The timing when ACE controlled the camera is different from the one when the camera-person controlled the camera.
- The zoom rate which ACE set is obviously different from the one which the camera-person set.

Table 3 shows the result of our comparing. In this table, "Case (a) $\sim$ (e)" are the reasons why we judged that those two scenes were "Not similar".

Case (a) ACE mistook image processing.

Case (b) ACE's zoom rate was different from the camera-person's one.

Case (c) The timing when ACE controlled the camera was too early.

Case (d) The timing when ACE controlled the camera was too late.

Case (e) The target area extracted by ACE was different from that of the camera-person.

The scene captured by ACE is similar to the one captured by the camera-person by a ratio of over 80%. ACE cannot capture, however, the similar scene due to some reasons. In the cases of (b), (c) and (d) in Table 3, ACE could capture similar scene although the timing of camera control or the zoom rate was different. These are not so big problem for a lecture scene. On the other hand, in the cases of (a) and (e), ACE obviously captured different and unsuitable scene for a lecture.

5.3 Evaluation by Teacher

We showed the teacher two video lectures. The teacher evaluated whether ACE or the camera-person could capture the scene which he wanted to capture. Table 4 shows the result of comparing. "Unsuitable" shots are categorized following five reasons;

Case (A) The target area wasn't put in center of the shot.

Case (B) The zoom rate was unsuitable.

Case (C) The blackboard side should be captured.

Case (D) The screen side should be captured.

Case (E) The screen side should be kept on capturing.

Camera-persons could take satisfactory shots by a ratio of about 90%. On the other hand, ACE could take satisfactory shots by a ratio of about 80%. The ratio of Case (C), (D), and (E) of ACE are higher than those of camera-persons. The periods of time which the teacher had selected as "Unsuitable" were little less than those of "Not similar" in Table 3.

6 Discussion

ACE could capture lecture scenes much better than the fixed camera's shots. Moreover, the scene captured by ACE is as good as the one captured by the camera-person. Some students said that the scene captured by ACE is not good because ACE captured the blackboard and the screen by using only one camera. For example, when ACE was capturing the blackboard side, students just saw the side even if they wanted to see the screen side, and vice versa. ACE was improved as compared with the previous version in the following points; of ACE.

- ACE came to be able to estimate teacher's behavior.
- ACE came to be able to extract the area explained by a teacher more robustly.
- ACE came to be able to capture better scenes.

7 Conclusion

We have designed a camera control strategy for capturing a lecture and developed a prototype of ACE. We have implemented our system by estimating teacher's behavior. New version of ACE can capture the lecture scene as well as camera-persons. We make sure that ACE is a useful tool for capturing a lecture.

ACE extracts an area explained by a teacher by only image processing. On the other hand, a camera-person extracts the area by observing and hearing the lecture. We will get ACE analyze the teacher's voice and recognize a keyword in the lecture. We guess that analyzing the teacher's voice helps ACE extracting more suitable explained object.

References

[1] A. Suganuma and S. Nishigori, Automatic Camera Control System for a Distant Lecture with Videoing a Normal Classroom, *Proc. World Conference on Educational Multimedia, Hypermedia & Telecommunications*, 2002, 1892–1897.

[2] A. Suganuma and T. Ashikawa, Automatic Camera Control System with Both Image and Sound Processing, *Proc. Computers and Advanced Technology in Education*, 2003, 722–727.

[3] http://www.sfc.wide.ad.jp/DVTS/index.html

[4] N. Johnson and D. Hogg, Representation and synthesis of behavior using Gaussian mixtures, *Proc. Image and Vision Computing 20*, 2002, 889–894.