International Journal of Computer Vision 37(2), 175–185, 2000

Integrated Person Tracking Using Stereo, Color, and Pattern Detection

T. DARRELL*, G. GORDON, M. HARVILLE AND J. WOODFILL
Interval Research Corp., 1801C Page Mill Road, Palo Alto CA 94304, USA

Abstract. We present an approach to real-time person tracking in crowded and/or unknown environments using integration of multiple visual modalities. We combine stereo, color, and face detection modules into a single robust system, and show an initial application in an interactive, face-responsive display. Dense, real-time stereo processing is used to isolate users from other objects and people in the background. Skin-hue classification identifies and tracks likely body parts within the silhouette of a user. Face pattern detection discriminates and localizes the face within the identified body parts. Faces and bodies of users are tracked over several temporal scales: short-term (user stays within the field of view), medium-term (user exits/reenters within minutes), and long term (user returns after hours or days). Short-term tracking is performed using simple region position and size correspondences, while medium and long-term tracking are based on statistics of user appearance. We discuss the failure modes of each individual module, describe our integration method, and report results with the complete system in trials with thousands of users.

Keywords: face detection, tracking and recognition, human-computer interface, frame-rate stereo, multi-modal integration

1. Introduction

The creation of displays or environments which passively observe and react to people is an exciting challenge for computer vision (Maes et al., 1996; Rehg et al., 1997). Faces and bodies are central to human communication and yet machines have been largely blind to their presence in real-time, unconstrained environments. Often, computer vision systems for person tracking exploit a single visual processing technique to locate and track user features. These systems can be non-robust to real-world conditions with multiple people and/or moving backgrounds. Additionally, tracking is usually performed only over a single, short time scale: a person model is typically based only on an unbroken sequence of user observations, and is reset when the user is occluded or leaves the scene temporarily.

We have created a visual person tracking system which achieves robust performance through the integration of multiple visual processing modalities and by temporal scales. With each modality alone it is possible to track a user under optimal conditions, but each also has, in our experience, substantial failure modes in unconstrained environments. Fortunately these failure modes are often independent, and by combining modules in simple ways we can build a system which is relatively robust.

We will show how our system works well in visually noisy environments and does not make any assumptions about static background patterns. Additionally we will also show how our system is robust to the failure of each individual module, and that adding a given module to the system always increases overall performance. A key strength of our system is the use of real-time stereo depth estimation hardware, described below; other authors have proposed systems for person tracking with multiple cues (for example see Toyama and Hager (1996), Regh et al. (1999) and Isard and Blake (1998)) but have not incorporated video-rate range as one of the processing modalities.

In the following sections we describe our tracking framework and the three vision processing modalities used. We then describe an initial application of our

*Present address: MIT, A1 Lab, 545 Technology Square, Cambridge, MH 02139, USA. Email: trevor@ai.mit.edu

system: a face-responsive, interactive video display. Finally we show the results of our system when deployed with naive users, and analyze both the qualitative success of the application and the quantitative performance of our tracking algorithms.

2. Tracking Framework

A person tracking system for interactive environments has several desired criteria: it should operate in real-time, be robust to multiple users and changing background, provide a relatively rich visual description of the users, and be able to track people when they are occluded or momentarily leave the scene. We achieve these goals through the use of multi-modal integration and multi-scale temporal tracking.

We base our system on three primary visual processing modules: depth estimation, color segmentation, and intensity pattern classification (see Fig. 1). As described in more detail below, depth information is estimated using a dense real-time stereo technique and allows easy segmentation of the user from other people and background objects. An intensity-invariant color classifier detects regions of flesh tone on the user and is used to identify likely body part regions such as face and hands. A face detection module is used to discriminate head regions from hands and other tracked body parts.

Figure 2 shows the output of the three vision processing modules. As a person tracker, each is individually fragile: head-sized objects (e.g. a notebook) can cause false positives in the range module, flesh-like colors found in other materials can cause false positives in the color module, and face pattern detectors typically are slower and cause false negatives in non-canonical poses or expressions. However, when integrated together these modules can yield robust, fast tracking performance.

Tracking is performed in our system on three different time-scales: short-range (frame to frame while the person is visible), medium-range (when the person is momentarily occluded or leaves the field of view for a few minutes), and long range (when the person is absent for hours, days or more.) Longer-term tracking can be thought of as a person identification task, where the database is formed from the set of previous users. For short-term tracking we simply compute region correspondences specific to each processing modality based on region position and size. Multi-modal integration is performed using the history of short-term tracked regions from each modality, yielding a representation of the user's body shape and face location.

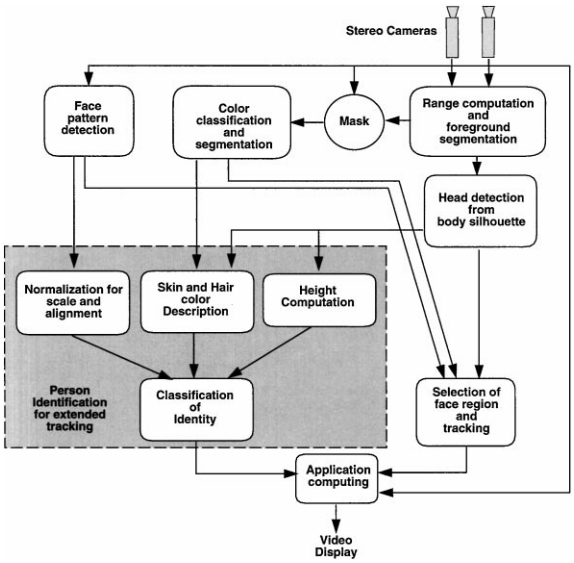


Figure 1. System overview showing the relationship of each modality with detection and short-term tracking, and with long-term tracking/identification.

We rely on a statistical model of multi-modal appearance to resolve correspondences between tracked users over time. In addition to body shape and face appearance location, the color of hair, skin, and clothes is recorded at each time step. We record the average value and covariance of represented features and use them to identify users when they return. For medium-term tracking lighting constancy and stable clothing color are assumed; for long-term tracking we adjust for changing lighting and do not include clothing in the match criteria.

We now discuss module specific processing, including classification, segmentation/grouping, and short-term tracking. Following that, we present our integration scheme, and correspondence method for medium and long-term tracking. Pixel-wise classification, grouping and short-term tracking are performed independently in each modality. Stereo processing outputs a user's silhouette defined by range regions, color processing yields a set of skin color regions within range silhouette boundaries, and face processing returns a list of detected frontal face patterns; we describe each module in turn.

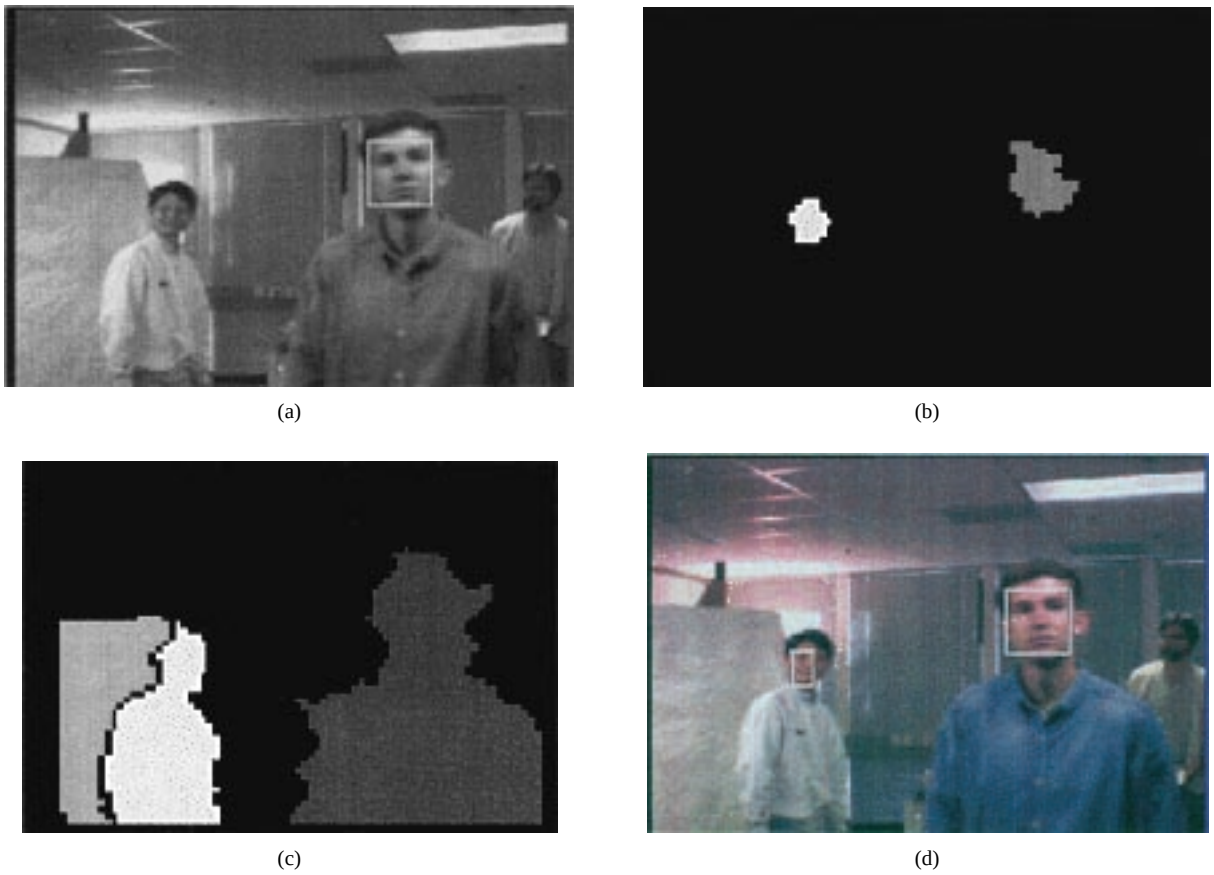


Figure 2. Output of vision processing modules: (a) face pattern detection output, (b) flesh hue regions from skin hue classification, (c) connected components recovered from stereo range data, and (d) shows the input image with boxes indicating the location of tracked users.

2.1. User Silhouette from Dense Stereo

To compute a set of user silhouettes, we rely on a dense real-time stereo system. Video from a pair of cameras is used to estimate dense range using the census stereo algorithm (Zabih and Woodfill, 1994); we have implemented this algorithm on a reconfigurable, FPGA based computing engine resident on a single PCI card. This stereo system searches a window of 32 possible discrete stereo disparities at each pixel on 320 by 240 images at over 50 frames per second (30 frames per second for standard video), or over 120 million pixel-disparities per second. These processing speeds compare favorably with other real-time stereo implementations such as (Kanade et al., 1996). With sub-pixel interpolation, eight bits of stereo depth information are available.

Our segmentation and grouping technique proceeds in several stages of processing, as illustrated in Fig. 3. We first smooth the raw range signal to reduce the effect of low confidence stereo disparities using a morphological closing operator. We then compute the response of a gradient operator on the smoothed range data and threshold at a critical value based on the largest magnitude depth discontinuity expected in the range profile of one person (e.g. approx 8 inches). Connected components analysis is applied to these regions of smoothly varying range. We return all connected components whose area exceeds a minimum threshold.

The range processing module provides these extracted user silhouettes, as well as estimates of head location. A candidate head is placed below the vertical maxima of the silhouette, in a mannar similar to Darrell et al. (1994) and Wren et al. (1997). Head position is refined in the integration stage, as described below.

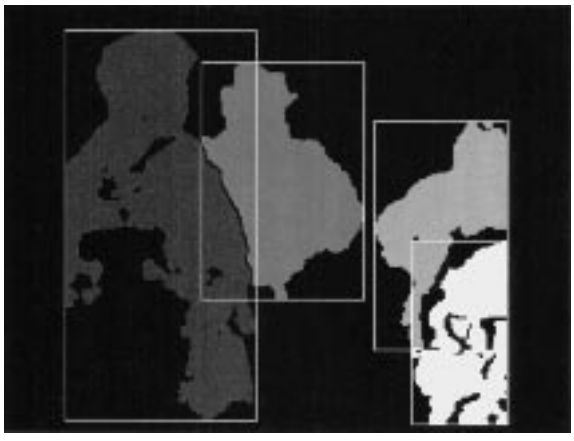
Disparity estimation, segmentation, and grouping are repeated independently at each time step; range silhouettes are tracked from frame to frame based on

(a)

(b)

(c)

(d)



(e)

Figure 3. Stereo range processing to extract user silhouettes. (a) left/right image pair, (b) raw disparity computed using Census algorithm, (c) disparity after morphological smoothing, (d) regions of slowly varying disparity, and (e) silhouettes recovered after connected components grouping.

position and size constancy. The centroid and size of each new range silhouette is compared to silhouettes from the previous time step. "Short-term" correspondences are indicated using a greedy algorithm starting with the closest unmatched region; for each new region the closest old region within a minimum threshold is marked as the correspondence match.

2.2. Skin Color Localization

Skin color is a useful cue for tracking people's faces and other body parts. We detect skin using a classification strategy which matches skin hue but is largely invariant to intensity or saturation, as this is robust to shading due to illumination and/or the absolute amount of skin pigment in a particular person.

We apply color processing to images obtained from one camera. Each image is initially represented with pixels corresponding to the red, green, and blue channels of the image, and is converted into a "log color-opponent" space. This space can directly represent the approximate hue of skin color, as well as its log intensity value. We convert (R, G, B) tuples into tuples of the form $(\log(G), \log(R) - \log(G), \log(B) - (\log(R) + \log(G))/2)$. Skin color is detected using a classifier with an empirically estimated Gaussian probability model of "skin" and "not-skin" in the log color-opponent color space. When a new pixel p is presented for classification, the likelihood ratio $P(p = skin)/P(p = non - skin)$ is computed as a classification score. Our color representation is similar to that used in (Fleck et al., 1996), but we estimate our classification criteria from examples rather than apply hand-tuned parameters. For computational efficiency at run-time, we may precompute a lookup table over all possible color values.

After the lookup table has been applied, segmentation and grouping analysis are performed on the classification score image. Similar to the range case, we use morphological smoothing, threshold above a critical value, and apply connected component computation. However, there is one difference: before smoothing we apply the low-gradient mask from the *range* modality. This restricts color regions to be grown only within the boundary of range regions; if spurious background skin hue is present in the background it will not adversely affect the shape of foreground skin color regions.

As with range processing, classification, segmentation, and grouping are repeated at each time step. Short-term tracking is performed on recovered color regions based on similar centroid position and region size. When a color region changes size dramatically, we check to see if two regions merged, or if one region split into two. If so we record the identity of the split or merged regions, to be used in the integration stage as described below.

Skin color regions that are above the midline of their associated range component, and which are appropriately sized at the given depth to be heads, are labeled as candidate heads and passed to the integration phase.

2.3. Face Pattern Detection

To distinguish head from hands and other body parts, and to localize the face within a region containing the head, we use pattern recognition methods which directly model the statistical appearance of faces based on intensity.

We based our implementation of this module on the CMU face detector (Rowley, 1996) library. This library implements a neural network which models the appearance of frontal faces in a scene, and is similar to the pattern recognition approach described in Poggio and Sung (1994). Both methods are trained on a structured set of examples of faces and non-faces.

Face detection is initially applied over the entire image; when one or more detections are recorded, they are passed directly as candidate head locations to the integration phase. Short term tracking is implemented by focusing search in a new frame within windows around the detected locations in the previous frame. If a new detection is found within such a window it is considered to be in short-term correspondence with the previous detection.

If no new detection is found and the detection in the previous frame overlapped a color or range region which was tracked successfully, then the face detection is updated to move with that region (as long as it persists and no new face detection is found). We record the relative offset of the face detection head with respect to the range or color head, and maintain that relationship in subsequent frames. This has the desired effect of allowing face detection to discriminate between head and hand regions in subsequent frames even when there may not be another face detection for several frames.

There is one special case in propagating face detection candidate heads. If the two color regions split or

merge as described above, we take steps to allow the face detection candidate head to follow the appropriate color region. We assume that the face is stationary between frames when deciding what color region to follow. If two regions have merged, the virtual detection follows the merged region, with offset such that the face's absolute position on the screen is the same as the previous frame. If two regions have split, the face follows the region closest to its position in the previous frame. These heuristics are simple, but work in many cases where users are intermittently touching their face with their hands.

2.4. Integrated Tracking

Our integration method is designed to take advantage of each module's strengths: range is typically fast but coarse, color is fast and prone to false positives, and face pattern detection is slow and requires canonical pose and expression. We place priority on face detection hits, when available, and use color or range to update position from frame to frame.

For each range silhouette, we collect the range, color, and face detection candidate head features. We compute the location of a user's head on the range silhouette as follows: if a face detection candidate head is present we return its location, otherwise we return any location with overlapping range and color candidates. If there is no overlap between candidates we use the location of the range candidate, or the location of a color candidate, in order of preference.

When the head location has been found, we update the estimate of head size. We recompute size based on the results of the face detector and the range modules. When a face detection result has been found, we use it to determine the real size of the face. If no face detection hit has been found, we use an average model of real face size.

Our system can be configured in two modes: single- or multiple-person tracking. Single-person mode is most appropriate for interactive games or kiosks which are restricted to a single user; multiple-person is more appropriate for general surveillance and monitoring applications. In single person mode, we return only a single range silhouette; we initially choose the closest range region, and then follow that region until it is no longer tracked in the short-term. In multiple-person mode each observed person is tracked simultaneously. The maximum number of people that can be tracked is limited by minimum size constants in the module specific connected components code.

3. Long-Term Tracking

When users are momentarily occluded or exit the scene, short-term tracking will fail since position and size correspondences in the individual modules are unavailable. To track users over medium and long-term time scales, we rely on statistical appearance models. Each visual processing module computes an estimate of certain user attributes, which are expected to be stable over longer time periods. These attributes are averaged as long as the underlying range silhouette continues to be short-term tracked, and are then used in a classification stage to establish medium and long-term correspondences.

Like multi-modal person detection and tracking, multi-modal person appearance classification is more robust than classification systems based on a single data modality. Height, color, and face pattern each offer independent classification data and are accompanied by similarly independent failure modes. Although face patterns are perhaps the most common data source for current passive person classification methods, it is unusual to incorporate height or color information in identification systems because they do not provide sufficient discrimination to justify their use alone. However, combined with each other and with face patterns, height and color can provide important cues to disambiguate otherwise similar people, or help classify people when only degraded data is available in other modes.

3.1. Observed Attributes

In the range module, we estimate the height of the user and use this as an attribute of identity. Height is obtained by computing the median value of the highest point of a user silhouette in 3-D. In the color module, we compute the average color of the skin and hair regions, and optionally a color histogram of clothing appearance. We define the hair region to be those pixels within a threshold distance to the top or sides of the head region; we only take those pixels which are of appropriate range and which are not classified as skin color. Clothing can be defined as all other pixels on the range silhouette which are not labeled as skin or hair.

In the face detector, we record an image of the actual face pattern wherever the detector records a hit. When a region is identified as a face based on the face pattern detection algorithm, the face pattern (greyscale subimage) in the target region is normalized and then passed to the classification stage. For optimal classification, we want the scale, alignment, and view of detected faces to be comparable. We resize the pattern to normalize for size, and discard images which are not in canonical pose or expression, which is determined by normalized correlation with an average canonical face.

For "medium-term" tracking, e.g., over seconds or minutes of occlusion or absence, we rely on all of the above attributes. For "long-term" tracking, over hours or longer, we cannot rely on attributes which are not invariant with time of day or from day to day: we correct all color values with a mean color shift to account for changing illumination, and exclude clothing color from the match computation.

3.2. *Classification*

In general, we compute statistics of these attributes while users are being tracked over short-term time periods, and compare against stored statistics of previously observed users to obtain medium- and long-term correspondences. User models can be acquired through an explicit training process with known identification strings, or by automatically instantiating models for each new short-term user track that can not be matched to a previous model.

When we observe a new person, we see if there is a previously tracked individual which could have generated the current observations. We find the previous individual most likely to have generated the new observations; if this probability is above a minimum threshold, we label the currently tracked region as being in correspondence with the previous individual. We integrate likelihood over time and modality: at time t, we find the identity estimate

$$u^* = \arg\max_j \; P(U_j \mid \omega) \tag{1}$$

where

$$P(U_j \mid \omega) = P(U_j \mid F_0, \ldots, F_t, H_0, \ldots, H_t, C_0, \ldots, C_t) \tag{2}$$

where F_i,H_i, and C_i are the face pattern, height, and color observations at time i, and U_j are the saved statistics for person j. We restart time at $t = 0$ when a new range silhouette is tracked. For the purposes of this discussion we assume $P(U_j)$ is uniform across users. With Bayes rule and the assumption of modality independence, we have:

$$u^* = \arg\max_j \left(P(F_0, \ldots, F_t \mid U_j) P(H_0, \ldots, H_t \mid U_j) P(C_0, \ldots, C_t \mid U_j)\right) \tag{3}$$

Since our observations are independent of the observed noise in our sensor and segmentation routines, the posterior probabilities at different times may be considered independent. With this we can incrementally compute probability in each modality:

$$P(F_0, \ldots, F_t \mid U_j) = P(F_0, \ldots, F_{t-1} \mid U_j) P(F_t \mid U_j) \tag{4}$$

and similarly for range and color data.

We collect mean and covariance data for the observed user color data, and mean and variance of user height; the likelihoods $P(F_i \mid U_j)$ and $P(C_i \mid U_j)$ are computed assuming a Gaussian density model. For face pattern data, we store the size- and position-normalized mean pattern for each user, and approximate $P(F_t \mid C_p)$ with an empirically determined density which is a function of the normalized correlation of F_t with the the mean pattern for person j.

Our present implementation does not have any time-bias—previously viewed individuals are equally likely to be recognized independent of the interval since their earlier interaction with the system. This is a reasonable assumption if no prior knowledge about visit time statistics is available for a particular application; if such knowledge is available then the inter-visit interval should be included as a data term when computing the liklihood of a model given the current observation.

4. A Real-Time Face-Responsive Display

Our initial application of our integrated, multi-modal visual person tracking framework is a face-responsive visual display. We construct a video display where cameras observe the user from the same optical axis as used by the display, and send estimates of the 3-D head position of observers of the screen to the application program.

We create a virtual mirror by placing cameras so that they share the same optical axis as a video display, using a half-silvered mirror to merge the two optical paths.

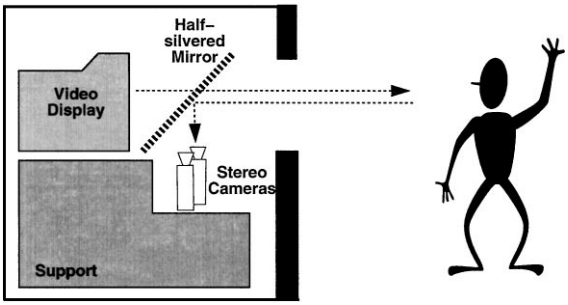


Figure 4. Display and viewing geometry for interactive video kiosk applications: cameras and video-display share optical axis through a half-silvered mirror.

The cameras view the user through a 45-degree half mirror, so that the user can view a video monitor while also looking straight into (but not seeing) the cameras. Video from one camera is displayed on the monitor after the application of various computer graphics distortion effects, so as to create a virtual mirror effect. Figure 4 shows the display and viewing geometry of our apparatus.

One application we have explored using this display is an interactive graphics experience. Users approach the display, and see video of their faces distorted with real-time special effects. The effect is a virtual funhouse mirror, but in which only the face regions are distorted. Using video texture mapping and the OpenGL graphics system, we have implemented graphics methods to distort faces on the screen using one of the following special effects: spherical expansion, spherical shrinking, swirl, lateral expansion, and a vertical melting effect. This creates a novel and entertaining interactive visual experience where users get immediate visual feedback from their tracked faces. Our tracking system is not limited to this type of interactive special effect, of course, but we have found it to be an intuitive and entertaining demonstration of the capabilities of the system.

Our system is currently implemented using three computer systems (one PC, two SGI O2), a large NTSC video monitor, stereo video cameras, a dedicated stereo computation PC board, and the half-mirror imaging apparatus. The full tracking system, including all vision and graphics processing, runs at approximately 12 Hz with approximately 100 ms delay.

5. Results

We have demonstrated versions of our system at several conferences and one public museum installation (Darrell et al., 1997, 1998). Conservatively, we estimate that over ten thousand users have experimented with our system. In the interactive video distortion application the goal of the system is to identify the 3-D position and size of users' heads in the scene, and apply a distortion effect in real-time only over the region of the image containing the user's face. The distorted image was then displayed on the virtual mirror screen. In the first version of the system (Darrell et al., 1997) only a single user was tracked and there was no long-term identification capability; the most recent version tracked all users present and implemented the long-term tracking described above.

Qualitatively, the system was a complete success. Our tracking system was able to localize video distortion on the user's face accurately enough for them to experience the desired perceptual effect. Overall, users reported that the system was interesting and entertaining. Figure 5 shows a typical final image displayed on the virtual mirror. The system performed well with both single users and crowded conditions, and despite environments which were often quite visually noisy. At several conference sites visual effects from other exhibits were randomly projected onto both the background and the people being tracked by our system; this would have caused great difficultly for systems which relied on static color background models rather than real-time stereo data for obtaining body silhouettes.

Figure 5. Example distortion output from interactive special-effect application.

5.1. Evaluation

We quantitatively evaluated the performance of our system using two different datasets: a set of stills captured at the conference installations to evaluate detection performance, and video sequences of users in our laboratory who interacted with the system over several days to evaluate medium- and long-term tracking.

At the conference installation, we sampled the performance of the system every 15 seconds over an evaluation period of approximately 3 hours. At each sample point we captured both a color image of the scene and an image of the output of the range module after disparity smoothing. We segregated image/range pairs of scenes with no users present, after verifying that our system did not generate any false positive detection on these images. We retained approximately 300 registered color/range pairs with one or more people present.

We also collected a similar set of approximately 200 registered range/color stills of users of the system while on display in our laboratory, similar to the images in Figs. 2 and 3(a). Table 1 summarizes the single-person detection results we obtained on these test images. A correct match was defined when the corners of the estimated face region were sufficiently close to manually entered ground truth (within $\frac{1}{4}$ of the face size). Overall, when all modules were functioning, we achieved a success rate in excess of 96%; when the color or face detection module was removed performance was still above 93%, indicating the power of the range cue for detecting likely head locations. Face detection results for the conference data set are not reported as we were not able to record images at sufficient resolution to fully evaluate detection performance off-line. We note that faces in our dataset were often quite small and our application encouraged unusual expression and pose; this explains the decreased performance relative to more traditional face detection databases.

Performance was generally better on the Conference dataset than on the Lab dataset, which we believe was due to the fact that users in the latter dataset were more familiar with the system and attempted to push the limits of the tracking system. Users in the Conference dataset exhibited fewer hand gestures and limb movements, allowing overly simple heuristics such as employed by the range module alone to accurately identify head location in a relatively large proportion of trails.

To evaluate our longer term tracking performance we used statistics gathered from 25 people in our laboratory who visited our display several times on different days. This population included multiple races, an even distribution of genders, and a relatively wide range of adult heights. People's hairstyle, clothing, and the exterior illumination conditions varied between the times data were collected. We tested whether our system was able to correctly identify users when they returned to the display. During this time period no other users interacted with the system. In general, our results were better for medium term tracking (intra-day) than for long term (inter-day) tracking, as would be expected. Table 2 shows the extended tracking results: the correct classification percentage is shown for each modality and for the combined observations from all modes. This table reflects the recognition rate using all of the data from each short-term tracking session: on average, users were tracked for 15 seconds before short-term tracking failed or they exited the workspace.

By integrating modes we were able to correctly establish correspondences between tracked users in all of the medium-term cases, which typically involved temporal gaps between 10 and 100 seconds. In the long-term cases, which typically reflected gaps of one day, integrated performance was 87%. Performance was

Table 1. Face detection and localization results on Conference and Lab datasets using different combinations of input modules, ordered by increasing error rate.

Modules enabled				
Color	Range	Pattern	Conference data	Lab data
√	√	√	97%	96%
	√	√	97%	95%
√	√		97%	93%
	√		97%	90%
√		√	92%	93%
√			90%	89%
		√	*	80%

Table 2. Extended tracking performance: correct identification rate at end of session.

Performance	Medium-term (intra-day)	Long-term (inter-day)
Height	44%	20%
Color	84%	63%
Face pattern	84%	67%
Multi-modal	100%	87%

generally worse across longer time intervals. While there are many unknown factors which contribute to this effect, we suspect that variations in scene illumination, user clothing and footware, and user behavior are larger inter-day rather than intra-day and can explain the observed performance. Further work is needed to fully analyze the various sources of classification uncertaintly across different time-scales.

A more complete depiction of medium- and long-term performance is shown in Fig. 6. These figures show the recognition rate vs rank threshold, i.e., the percentage of time the correct person was above a given rank in the ordered likelihood list of predicted users. As expected, identification becomes more reliable over time as more data is collected. Figure 7 shows the rank of the correct person over time, averaged across all test sessions; correct identification (average rank equals one) is almost always achieved within one second in the medium-term case, and within three seconds in the long-term case.

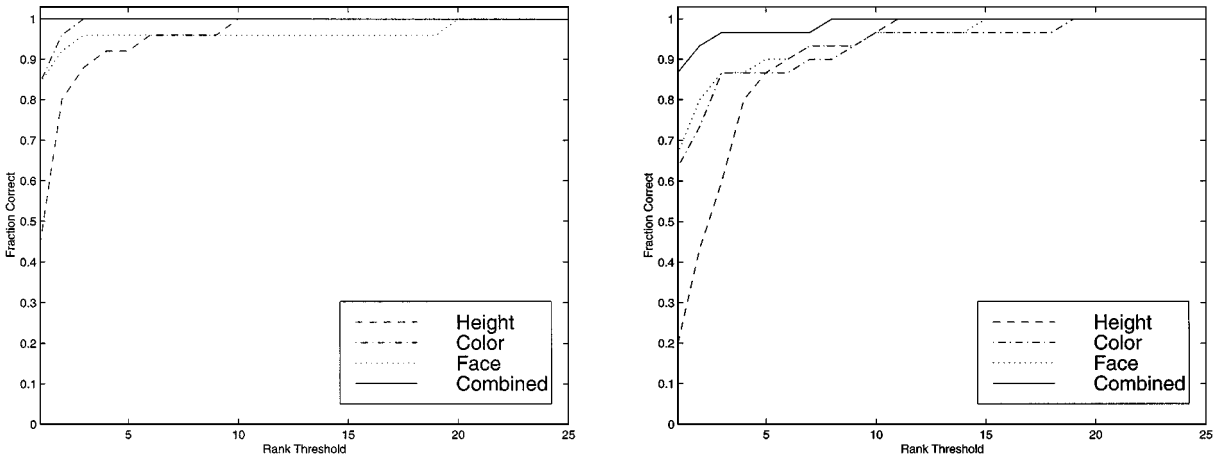


Figure 6. Identification performance for different thresholds on the size of the target set returned (the rank threshold). The highest probability models are returned in the target set. If the actual person is in the target set, the match is considered correct. Performance is shown for: (a) medium-term tracking, (b) long-term tracking. The left edge of the graph thus shows exact-match performance.

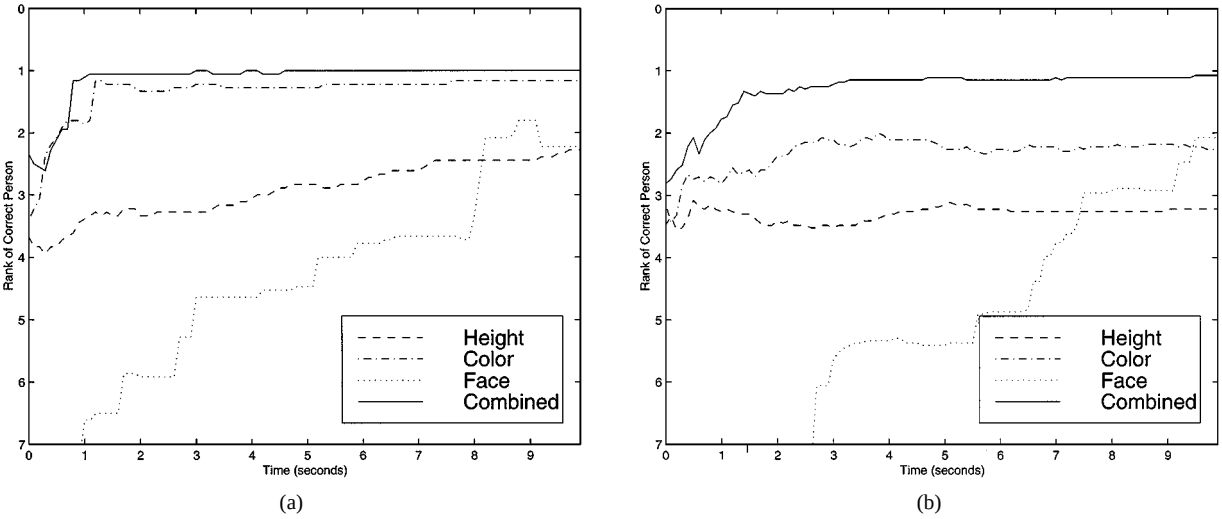


Figure 7. Average rank of correct person vs. time. (a) medium-term tracking, (b) long-term tracking.

5.2. Discussion

We draw two main conclusions from the detection results; first, that range data is a powerful cue to detecting heads in complex scenes. Second, integration is useful: in almost every case, the addition of modules improved system performance. Performance was generally high, but individual module results varied considerably across datasets. In particular the face pattern module fared relatively poorly on the Conference dataset. We believe that this is largely due to the small size and poor illumination of many of the faces in these images. Additionally, in both datasets our application encouraged people to make exaggerated expressions, which was beyond the scope of the training for this module.

In contrast, for extended tracking it is clear from these results that the face pattern is the most valuable of the three modes when we consider all the data available during the session. Face pattern data is most discriminating at the *end* of the test session; however, the other modalities are dominant early in the session. The face detection module operates more slowly than the other modes, so the face pattern data is not available immediately and accumulates at a slower rate. Therefore, in the first few seconds the overall performance of the extended tracking system is due primarily to color and height data, and far exceeds the performance based on face pattern alone.

6. Conclusion

We have demonstrated a system which can respond to a user's face in real-time using completely passive and non-invasive techniques. Robust performance is achieved through the integration of three key modules: depth estimation to eliminate background effects, color classification for fast tracking, and pattern detection to discriminate the face from other body parts. We use descriptions of the user computed from the same modalities to track over longer time scales when the user is occluded or leaves the scene. Our system has application in interactive entertainment, telepresence/virtual environments, and intelligent kiosks which respond selectively according to the presence, pose, and identity of a user. We hope these and related techniques can eventually balance the I/O bandwidth between typical users and computer systems, so that they can control complicated virtual graphics objects and agents directly with their own expression.

References

Darrell, T., Gordon, G., Woodfill, W., and Baker, H. 1997. A magic morphin mirror. In *SIGGRAPH '97 Visual Proceedings*. ACM Press.

Darrell, T., Harville, M., Gordon, G., and Woodfill, W. 1998. Mass hallucinations. *SIGGRAPH '98 Visual Proceedings*. ACM Press. Also shown at CVPR'98 Demonstration Session, Santa Barbera, CA, June 1998, and at The Tech Museum of Innovation, San Jose, Oct. 1998–April 1999.

Darrell, T., Maes, P., Blumberg, B., and Pentland, A. 1994. A novel environment for situated vision and behavior. In *IEEE Workshop on Visual Behaviors, CVPR '94*, Seattle. IEEE CS Press.

Fleck, M., Forsyth, D., and Bregler, C. 1996. Finding naked people. In *European Conference on Computer Vision*, Vol. II, pp. 592–602.

Isard, M. and Blake, A. 1998. Condensation: Unifying low-level and high-level tracking in a stochastic framework. In *Proc. 5th European Conf. Computer Vision*, Vol. 1, pp. 893–908.

Kanade, T., Yoshida, A., Oda, K., Kano, H., and Tanaka, M. 1996. A video-rate stereo machine and its new applications. In *Computer Vision and Pattern Recognition Conference*, San Francisco, CA.

Maes, P., Darrell, T., Blumberg, B., and Pentland, A.P. 1996. The ALIVE system: Wireless, full-body, interaction with autonomous agents. *ACM Multimedia Systems*: Special Issue on on Multimedia and Multisensory Virtual Worlds.

Poggio, T. and Sung, K.K. 1994. Example-based learning for view-based human face detection. In *Proceedings of the ARPA IU Workshop '94*, Vol. II, pp. 843–850.

Rehg, J., Loughlin, M., and Waters, K. 1997. Vision for a smart Kiosk. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition, CVPR-97*. IEEE Computer Society Press, pp. 690–696.

Rehg, J., Murphy, K., and Fieguth, P. 1999. Vision-based speaker detection using Bayesian networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition, CVPR-99*. IEEE Computer Society Press, pp. 110–116.

Rowley, H., Baluja, S., and Kanade, T. 1996. Neural network-based face detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition, CVPR-96*. IEEE Computer Society Press, pp. 203–207.

Toyama, K. and Hager, G. 1996. Incremental focus of attention for robust visual tracking. In *Proceedings of the 1996 IEEE Conference on Computer Vision and Pattern Recognition*; pp. 189–195.

Woodfill, J. and Von Herzen, B. 1997. Real-time stereo vision on the PARTS reconfigurable computer. In *Proceedings IEEE Symposium on Field-Programmable Custom Computing Machines*, Napa, pp. 242–250.

Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A. 1997. Pfinder: Real-time tracking of the human body. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zabih, R. and Woodfill, J. 1994. Non-parametric local transforms for computing visual correspondence. In *Proceedings of the Third European Conference on Computer Vision*, Stockholm, pp. 151–158.