Robust Visual Person Tracking for Interactive Displays.

T. Darrell, G. Gordon, M. Harville, J. Woodfill
Interval Research Corp.
1801C Page Mill Road
Palo Alto CA 94304
trevor,gaile,harville,woodfill@interval.com

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. This is a short version of the report in [1].

1. Introduction

The creation of displays or environments which passively observe and react to people is an exciting challenge for computer vision [5, 7]. Faces and bodies are central to human communication and yet machines have been largely blind to their presence in real-time, unconstrained environments.

As reported in [1], we have created a visual person tracking system which achieves robust performance through the integration of multiple visual processing modalities and by tracking over multiple 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 with overall robust performance.

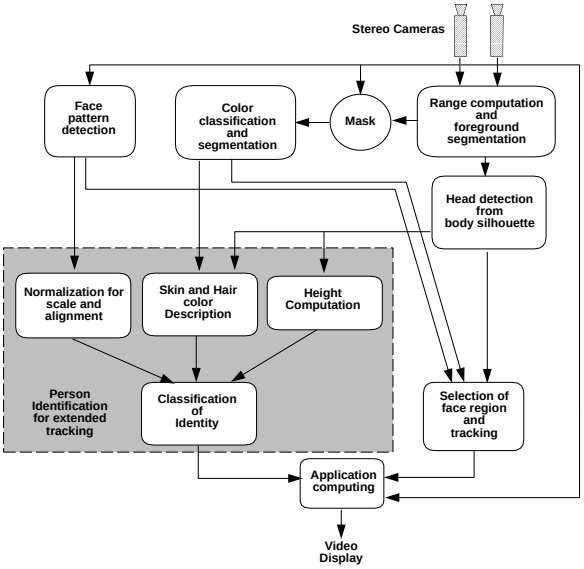

Figure 1. System overview showing the relationship of each modality with detection and short-term tracking, and with long-term tracking/identification.

In the following sections we describe our tracking framework and the three vision processing modalities used. We then describe an initial application of our system: a face-responsive, interactive video display.

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 inten-

sity pattern classification (see Figure 1). 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.

As a person tracker, each module is individually fragile: notebooks are indistinguishable from faces in range silhouette, flesh color signs or clothes fool color-only trackers, and face pattern detectors typically are slower and only work with relatively canonical poses and 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.) Long-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. Multimodal 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.

For medium and long-range tracking, we rely on a statistical model of multi-modal appearance to resolve correspondences between tracked users. In addition to body shape and face 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 for matching. 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.

3. Mode-specific processing

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. Each mode also provides an independent estimate of head location and performs short-term tracking.

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 a technique based on the census transform [9]; we have implemented the census algorithm on a single PCI card, using a multi-FPGA reconfigurable computing engine [10]. This stereo system is capable of computing 24 stereo disparities on 320 by 240 images at 42 frames per second, or approximately 77 million pixel-disparities per second. These processing speeds compare favorably with other real-time stereo implementations such as [4].

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 segmentation processing to images obtained from one camera of the stereo pair. 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 [3], but we estimate our classification criteria from examples rather than apply hand-tuned parameters. For computational efficiency at run-time, we precompute a lookup table over all possible color values.

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 [8] 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 [6]. 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, then the face detection is updated to move with that region (as long as it persists).

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. As described above, when a candidate pattern detection head overlaps with a range or color candidate head, it persists and follows the range or color region. 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.

For each frame, 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 it; otherwise we return any location with overlapping range and color candidates, the location of the range candidate, or the location of a color candidate, in order of preference.

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 virtual 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 it's position in the previous frame. These heuristics are simple, but work in many cases where users are intermittently touching their face with their hands.

When the head location has been found, we update the estimate of head size. We have found that color is a relatively unreliable estimator of size; instead, 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.

5. 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 tracked in the short-term, and 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.

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 would exclude clothing color from the match computation.

6. A Real-time Virtual Mirror Display

Our initial application of our integrated, multi-modal visual person tracking framework is to create 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. One application we have explored using this display is an interactive graphics experience in which users' faces are distorted in real-time. The effect is a virtual fun-house mirror, but in which only the face regions are distorted.

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. 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. 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 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 12Hz.

7. 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

[1] Darrell, T., Gordon, G., Harville, M., Woodfill, W., "Integrated person tracking using stereo, color, and pattern detection", Proceedings IEEE Conference Computer Vision and Pattern Recognition (CVPR-98), pp. 601-609, Santa Barbera, 1998. See also http://www.interval.com/papers/1998-021.

[2] Darrell, T., Gordon, G., Woodfill, W., Baker, H., A Magic Morphin Mirror, SIGGRAPH '97 Visual Proceedings, ACM Press. 1997.

[3] Margaret Fleck, David Forsyth, and Chris Bregler (1996) "Finding Naked People," European Conference on Computer Vision , Volume II, pp. 592-602. 1996.

[4] Kanade, T., Yoshida, A., Oda, K., Kano, H., and Tanaka, M., "A Video-Rate Stereo Machine and Its New Applications", Computer Vision and Pattern Recognition Conference, San Francisco, CA, 1996.

[5] Maes, P., Darrell, T., Blumberg, B., and Pentland, A.P., "The ALIVE System: Wireless, Full-Body, Interaction with Autonomous Agents". ACM Multimedia Systems: Special Issue on on Multimedia and Multisensory Virtual Worlds, Sprint 1996.

[6] Poggio, T., Sung, K.K., "Example-based learning for view-based human face detection". Proceedings of the ARPA IU Workshop '94, II:843-850. 1994.

[7] Rehg, J., Loughlin, M., and Waters, K., "Vision for a Smart Kiosk", Proc. IEEE Conf. Computer Vision and Pattern Recognition, CVPR-97, pp. 690-696. IEEE Computer Society Press. 1997.

[8] Rowley, H., Baluja, S., and Kanade, T., "Neural Network-Based Face Detection", Proc. IEEE Conf. Computer Vision and Pattern Recognition, CVPR-96, pp. 203-207,. IEEE Computer Society Press. 1996.

[9] Zabih, R., and Woodfill, J., "Non-parametric Local Transforms for Computing Visual Correspondence", Proceedings of the third European Conference on Computer Vision, Stockholm, pp. 151 - 158. May 1994.

[10] Woodfill, J., and Von Herzen, B., "Real-Time Stereo Vision on the PARTS Reconfigurable Computer", Proceedings IEEE Symposium on Field-Programmable Custom Computing Machines', Napa, pp. 242-250, April 1997.