Bayesian Modality Fusion: Probabilistic Integration of Multiple Vision Algorithms for Head Tracking

Kentaro Toyama **Eric Horvitz**
Microsoft Research
Redmond, WA 98052-6399
{**kentoy,horvitz**}**@microsoft.com**

ABSTRACT

We describe a head-tracking system that harnesses Bayesian modality fusion, a technique for integrating the analyses of multiple visual tracking algorithms within a probabilistic framework. At the heart of the approach is a Bayesian network model that includes random variables that serve as context-sensitive indicators of reliability of the different tracking algorithms. Parameters of the Bayesian model are learned from data in an offline training phase using ground-truth data from a Polhemus tracking device. In our implementation for a real-time head tracking task, algorithms centering on color, motion, and background subtraction modalities are fused into a single estimate of head position in an image. Results demonstrate the effectiveness of Bayesian modality fusion in environments undergoing a variety of visual perturbances.

1 INTRODUCTION

Despite intensive research efforts over the last decade, robust, vision-based head tracking remains difficult. Real-time tracking systems are often confused by waving hands or changing illumination. Face detection systems [22, 23] are seldom run at camera frame rates, and are limited to the analysis of frontal views of the face under controlled lighting conditions. A body of research suggests that no single visual modality is at once consistent enough to detect all heads and yet discriminating enough to detect heads only. Color, for example, changes with shifts in illumination. On the other hand, "skin-color" is not restricted to skin.

In the robotics and target tracking communities, researchers have investigated a variety of *sensor fusion* techniques to unify the results of sets of sensors [1, 19]. Of course, different types of data present in images, such as color, edge, and motion, can be considered different sensing modalities, and so fusion techniques can apply to wholly vision-based tracking, as well. Work that uses variations of the *Probabilistic Data Association Filter* [2] combines color and edge data for tracking a variety of objects [21]. Other approaches use color information as a prior to bias estimation based on edge data within a multiple-hypothesis framework [16].

More recently, research on real-time head tracking has focused on the combination of multiple visual cues. One approach considers edge and color data. Head position estimates are made by comparing match scores based on image gradients and color histograms. The estimate deemed most reliable is returned as the position [3]. Another approach heuristically integrates color data, range data, and frontal face detection for tracking [8]. A recent system based on this methodology is capable of tracking multiple heads simultaneously. This work highlights the potential for leveraging multiple modalities to enhance the robustness of head tracking. However, the methodology employed for fusing the different modalities is *ad hoc*, relying on the manual tuning of parameters.

We have pursued research on *Bayesian modality fusion* for head tracking. Our approach is motivated by the observation that the performance of any particular head-tracking algorithm may be satisfactory in some visual contexts but may degrade significantly in others—and that each algorithm has its own profile of sensitivity to context. We have sought to build a head-tracking system that overlays the results of multiple modalities in a coherent manner by computing the *context-sensitive reliability* of each modality and using these reliabilities to mesh the results into a single estimate of position. This approach is distinct from earlier work on head tracking with multiple modalities in its focus on incorporating context-sensitive evidence *about* reliability for integrating different sensing modalities, and in integrating the analyses of the modalities in a probabilistically coherent manner.

The use of Bayesian modality fusion for head tracking is related conceptually to prior research on the use of probabilistic sensor error submodels within Bayesian networks for performing diagnosis from real-time telemetry [14]. In that work, dynamically updated sensor-error models are used to interpret the relevance of sensor readings for diagnosing potential problems with the propulsion systems of the Space Shuttle. The sensor error models take into consideration observations that provide probabilistic evidence *about* the current reliability of sensors. The inferred reliabilities are considered in the fusion of multiple observations in a sound manner during automated diagnosis and decision support. We have applied this approach to the problem of integrating the results of a set of head tracking modalities.

We shall first overview principles of Bayesian modality fusion in Section 2. Then, in Sections 3 and 4, we describe the vision primitives of our real-time head tracking system and the details of data collection and training. In Sections 5 and 6, we review results demonstrating the effectiveness of

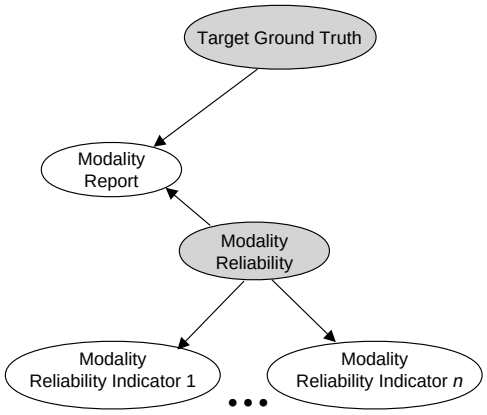

Figure 1. Bayesian network for inferring the ground truth about a visual target, conditioned on information about the report from a from a single modality.

Bayesian modality fusion for head tracking under a variety of visual conditions.

2 BAYESIAN MODELS FOR MODALITY FUSION

We harness Bayesian networks to capture probabilistic dependencies between the true state of the object being tracked (the *target*) and evidence obtained from tracking modalities. A Bayesian network is a directed acyclic graph that represents the joint probability distribution for a set of random variables [15, 17, 20]. Nodes in Bayesian networks represent random variables and arcs represent probabilistic dependencies among pairs of variables. The dependencies among variables in Bayesian network models can represent causal influences among variables.

Over the last decade, there have been significant strides in methods for constructing, learning, and performing inference with Bayesian-network models (see [11] for details of work in this community). Research has included the development of exact and approximate algorithms for Bayesian-network inference procedures [15], methods that allow for the induction of network structure from data [5, 13], and networks for reasoning over time [4, 6, 18]. Researchers have also examined conceptual links between Bayesian networks and probabilistic time-series analysis tools such as hidden Markov models (HMMs) and Kalman filters [6]. HMMs and Kalman filters can be represented by Bayesian networks with specific prototypical independencies and repetitive structure over time.

In constructing a Bayesian network for head tracking, we represent the true location of a user's head as a random variable. This "ground truth" variable influences other random variables, including *evidential* variables that are observed during tracking as well as intermediate, non-observed variables. Directed arcs among these variables capture probablistic influences. At run-time, the evidential variables are set to values that correspond to observed visual features—or the results of an analysis—and probabilistic inference is performed to compute a probability distribution over the true location.

In Bayesian models for modality fusion, we represent the output of distinct visual processing modalities as evidential variables. Additionally, we introduce special intermediate and evidential variables and dependencies that endow the model with the ability to perform real-time inference about the *context-sensitive reliabilities* of the different modalities.

Figure 1 displays the basic Bayesian network for Bayesian modality fusion. The nodes of the graph represent variables of interest, where the white nodes indicate variables that are instantiated by the vision modules and the gray nodes represent inferred values. The node labeled "Target Ground Truth" ($\mathbf{t}$) represents the unknown state of the target, and for our purposes, the overall goal of inference. From a Bayesian perspective, the ground-truth state causes the output from a visual modality (note that "causation" as used here comprises both deterministic and stochastic components). We indicate this influence with an arc from the ground truth to the node labeled "Modality Report" ($\mathbf{m}$).

The modality report is influenced also by its reliability, or its ability to accurately estimate ground-truth state ("Modality Reliability," ($\mathbf{r}$) – henceforth, referred to as *reliability*). Reliabilities themselves cannot be directly observed. The key notion behind Bayesian modality fusion is that both the reliabilities and the estimates of reliabilities vary with the structure of the scene being analyzed. To build a coherent framework for fusing reports from multiple modalities, we consider reliability as a variable and build probabilistic submodels to dynamically infer it as a function of easily ascertainable static or dynamic features of the image. In Figure 1, such evidence is represented by the n nodes labeled "Modality Reliability Indicator," ($\mathbf{i}_i$, *reliability indicator*) which are in turn influenced by the actual "Modality Reliability."

At run-time, the models for Bayesian modality fusion are instantiated with a set of observations including the modality report and the status of reliability indicators. The reliability of the modality report is computed and the inferred reliability and report are considered in inferring a probability distribution over the ground-truth state of the target object.

So far, we have considered a model for inferring the probability distribution over the true state of a target from a report by a single modality. We shall now generalize the model to represent multiple modalities. Figure 2 displays such a generalization. The diagram in the lower half of this figure shows a model with two modalities. Each modality includes a reliability submodel. We can include m different modalities in a similar manner.

Beyond the generalization to multiple modalities, Figure 2 also displays an extension of the representation to consider the status of variables at different times. Representing the temporal dynamics of a scene can provide valuable patterns of evidence for tracking algorithms and models of reliability.

As indicated in Figure 2, beyond employing variables that capture temporality via event definitions, we can build and assess models that consider the status of instances of variables at different periods of time. With this representation of time, the Bayesian network model is extended so that ensembles of variables are labeled with times. These

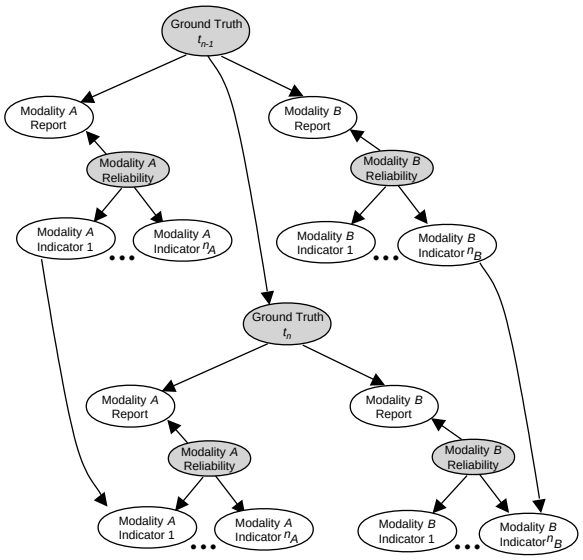

Figure 2. A dynamic Bayesian network model for integrating multiple visual processing modalities over time.

models capture dependencies between variables at different time periods, as well as among variables within a time slice. Representations of Bayesian networks over time that include temporal dependencies among some subset of variables have been referred to as *dynamic network models* and *dynamic Bayesian networks* [4, 7, 6, 18].

Figure 2 illustrates a Markov dynamic network model where the previous true state directly influences the current true state and where prior reliability indicators influence current indicators. If we were to simplify this model by assuming a single visual tracking mode, fixed modal reliabilities, and conditional influences in the form of linearly added Gaussian noise, this model reduces to a standard Kalman filter [6, 12]. By considering multiple modalities, modeling the details of probabilistic dependence, and considering the changing reliabilities of reports, we gain a flexible filter which weights estimates to different degrees based on their inferred accuracies.

3 BAYESIAN MODALITY FUSION IN HEAD TRACKING

We shall now describe details of learning and inference in the use of Bayesian modality fusion for head tracking. To perform tracking in real-time and to illustrate the effectiveness of the fusion algorithm, we will harness simple, computationally inexpensive algorithms for each of the visual processing components. We are not advocating the use of any specific modality; in many cases, more robust versions of the algorithms are widely known. Nonetheless, details of the vision algorithms are presented here to make our example concrete.

We implemented three visual modalities and identified a set of reliability indicators for each modality. The three modalities are (1) peak finding based on background subtraction, (2) color-based "blob" tracking, and (3) motion-based ellipse tracking. Each of these modes reports 4 values for the bounding box of the head (in image pixels) and 2 reliability indicators whose output types vary. For all three modalities, computation takes place on low resolution (80x60), subsampled images.

Background Subtraction Modality

Our background subtraction modality combines background subtraction with a peak detection scheme. Thresholding the difference between the current image and a stored background image immediately identifies foreground pixels if the camera is stationary. To accommodate deviations from this assumption, the stored background is updated in a manner similar to that described in [26].

Given a background image, $I_b(\cdot)$, we can easily determine *foreground* pixels as follows:

$$I_f(\mathbf{x},t) = \begin{cases} 1, \text{ if } |I(\mathbf{x},t) - I_b(\mathbf{x})| > k_f^{thresh} \\ 0, \text{ otherwise.} \end{cases}$$

After obtaining a foreground image, we then "drape" a horizontal line of points connected to their neighbors by spring forces onto the resulting image until the points hit significant clusters of foreground pixels [25]. Peaks in the draped line can be identified and the peak with the width and height closest to the previously known dimensions of the head are returned as the output.

Given a context of tracking a single person, we found that the accuracy of the draping strategy is correlated with the number of salient peaks detected by the system; multiple salient peaks is a strong indicator that the modality may be in a visual regime where it will not be an accurate predictor. We also noticed that the accuracy of the draping approach drops as the percentage of the image considered to be foreground rises or falls significantly, whether from a jolted camera or from the absence of foreground at all.

Therefore, we use two reliability indicators for the background subtraction modality and defined discretized variables to detect the status of reliability evidence at run-time. Reliability indicators for this method include the number of salient peaks in the draped line and the percentage of the image classified as foreground pixels.

Color-Based Tracking Modality

Color is an easily computed cue for head tracking. Various skin colors under likely illuminations can be approximated by a truncated pyramidal region in RGB space bounded by upper and lower thresholds on the ratios between red (r) and green (g) pixels, red and blue (b) pixels, and pixel intensity:

$$\begin{array}{ccccc} k_{rg}^{-} & < & r/g & < & k_{rg}^{+}, \\ k_{rb}^{-} & < & r/b & < & k_{rb}^{+}, \\ k_{int}^{-} & < & \frac{r+g+b}{3} & < & k+_{int}. \end{array}$$

In our color-based tracking modality, binary skin-color classification is performed over the entire image. Then, clusters of skin-colored pixels are identified by radiating investigative spokes outward from a skin-colored seed pixel until they hit non-skin-colored pixels [24]. The bounding box of the

cluster whose centroid and size are closest to the previous estimate is reported.

Using reasoning similar to that used for the background-subtraction modality, we defined reliability indicators for the the color-blob estimate as variables representing the status of the aspect ratio of the blob bounding box and the fraction of skin-colored pixels in the image.

Motion-Based Tracking Modality

Motion can also be a good indicator of head location, as people rarely hold their heads completely still. Pixels exhibiting motion can be detected by thresholding the difference between temporally adjacent image frames. In our motion-based modality, we set all motion-detected pixels to a constant, k_m. All other pixels experience a linear decay so that the final *decayed motion intensity* of the pixel at $\mathbf{x}$ is defined as follows:

$$I_m(\mathbf{x}, t_i) = \begin{cases} k_m, \text{ if } |I(\mathbf{x}, t_i) - I(\mathbf{x}, t_{i-1})| > k_m^{thresh}, \\ \max(0, I_m(\mathbf{x}, t_{i-1}) - 1), \text{ otherwise.} \end{cases}$$

Ellipse tracking is then performed on the motion intensity image by using conjugate gradient descent on the ellipse parameters to maximize the normalized sum of the motion intensity values lying beneath the ellipse (similar to [3]). In the approach, we fix the aspect ratio and consider position and scale over a range immediately surrounding the last known parameters.

Motion decay has been used before for "stateless" action recognition [9]. We use motion decay in the motion-based modality for tracking, given two desirable properties. First, the decay accumulates motion from previous frames, implicitly smoothing the motion image. Second, the decay creates a gradient in the motion image that rises with recency of motion. Thus, we can constrain the search range for ellipse tracking while maintaining robustness in the absence of motion filters—which often fail under jerky motion. As with the color-based modality, the bounding box of the final ellipse is used as the head position estimate from motion.

We defined reliability indicators for the motion-based modality in terms of the percentage of the motion identified in the image at hand and the residual of motion intensity observed under the final ellipse.

4 LEARNING MODEL PARAMETERS

To build a real-time model of Bayesian modality fusion, we need to populate the conditional probability tables defined by the structure of the Bayesian network for modality fusion. Such tables can be assessed with expert knowledge or learned through a training procedure. We designed a training system for acquiring the conditional probability tables by integrating a Polhemus Fastrak position-sensing device with the output from the three vision modalities. We took the Polhemus reading as ground-truth position. The Polhemus device was attached to the top of a subject's head, so that the center of the head in the horizontal plane was accurately determined as long as the subject's head remained upright.

For training purposes, Bayesian networks can be viewed as a set of conditional probability tables. These were populated by counting (and normalizing) the occurrence of joint events in the training data. For example, given all training instances when the position reported by the Polhemus sensor is $\bar{\mathbf{x}}$ and the reliability for a modality is $\bar{\mathbf{r}}$, we can compute the likelihood that the modality issues a report $\tilde{\mathbf{x}}$ (and that its indicators assume some values), simply by counting the times when this event occurs in the training data and dividing by the total number of events with $\bar{\mathbf{x}}$ and $\bar{\mathbf{r}}$. Additionally, we provided a prior on the network parameters, which act as "default values," when training data is lacking. The effect of the prior decreases with greater amounts of training data.

The conditional probability tables of the Bayesian network were populated by converting the training data into sets of probabilities representing the respective conditional contexts (*e.g.*, the probability of seeing specific values of a position by the color blob given the position reported by the Polhemus system, and given the width and aspect of the blob bounding box).

5 ILLUSTRATION OF REAL-TIME MODALITY FUSION

Once trained, the Bayesian network for modality fusion can be used to perform inference about head position given a set of real-time observations, including the reports generated by each of the vision processing modalities and their reliability indicators. More specifically, the model infers a probability distribution over position that would be reported by the Polhemus device given the findings.

We now examine the qualitative performance of Bayesian modality fusion for the methods described in Section 3. For purposes of illustration, we coarsen the discretization of all variables, eliminate temporal dependencies, and show results solely for horizontal position. The photos and inference results were generated with MSBN, a Bayesian network modeling and inference environment developed at Microsoft Research [10].

Figures 3 and 4 show the structure of the Bayesian networks used for our experiments. The attached bar graphs indicate the probability distributions over the states of each variable. Modality reports and ground truth are in pixels quantized to bins representing 40 pixels each. Reliabilities range from 0 to 40 and above, where smaller values represent greater expected accuracies. At run time, observational variables (white-filled nodes), are set to specific values by the tracking system and inference is performed to compute probability distributions over the states of the hypothesis variables (gray-fill), including the ground truth and reliabilities.

The cases displayed in the figures highlight the role of context-sensitive changes in the reliabilities of the different modalities. Both of the cases consider an identical (though permuted) set of reports from each of the modalities. However, the evidence about reliabilities changes, and, as a result, we see a shift in the modality that is weighted most heavily in the overall report of head position. In Figure 3, we see that the report from motion-based ellipse tracking (on the right) dominates the final estimate because the network infers that its reliability is high. The reliability itself was computed from its two child nodes whose values are observed directly (and hence concentrated in single bins).

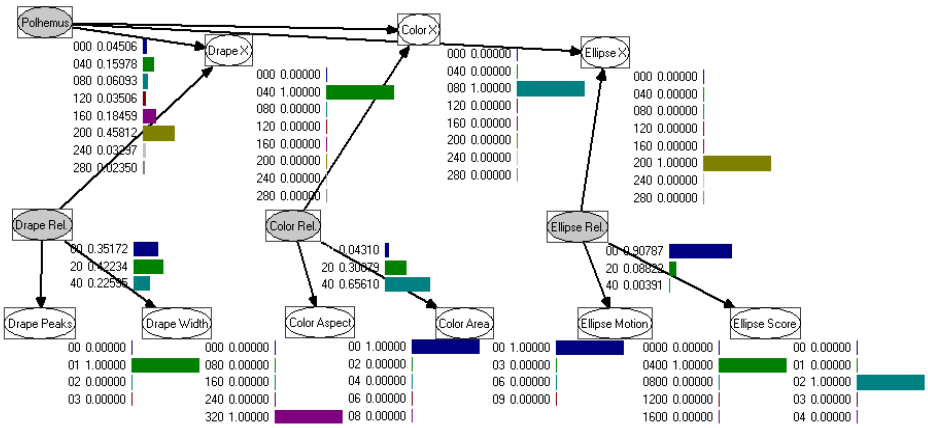

Figure 3. An estimate of head position dominated by the motion-based ellipse report. Color versions of all figures are available at http://research.microsoft.com/toyama/accv.pdf.

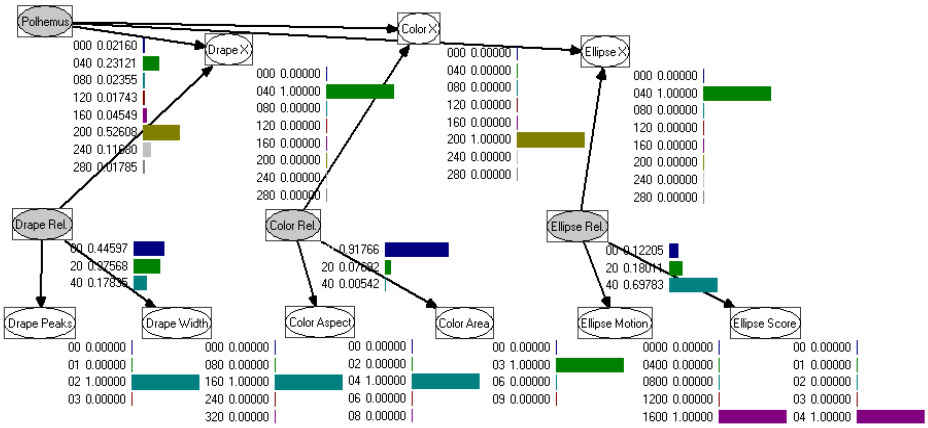

Figure 4. An estimate of head position dominated by the color-based blob report.

In Figure 4, that status of reliability indicators depress the inferred motion-based ellipse reliability and raise the color-based reliability, resulting in a final estimate that reflects the color-based report most strongly.

6 RESULTS

We now look at a system implemented on a 266MHz, single-processor Pentium II PC equipped with a Matrox Meteor framegrabber and a color camera. Conditional probabilities for the network were learned from empirical training data described in Section 4. The training data consisted of 10 minutes of data sampled at 10Hz. During the training session, illumination was varied, a person walked into the background, and the primary subjects intentionally presented signficant variations in position and pose.

All states and confidence values were discretized (at pixel granularity). State estimates with maximum probability were output as the final estimate. Execution cycled through the three modalities, with only one modality operating for any frame. Estimates of head position, however, were updated at frame rate (30Hz).

Over the course of a one-hour period in which 3 people entered and exited the field of view (to perform simple PC chores), the system correctly recognized and tracked heads for over 99% of the time that a person was in view (a total of 48 minutes). Unfortunately, this accuracy is not compelling as the subjects and the surrounding environment for this experiment were visually "well-behaved."

In a more interesting analyses, we explored success and failure modes during another set of experiments in which we deliberately caused visual perturbations. Figures 5 and 6 show the qualitative behavior of the algorithm during instances of tracking success. Under most circumstances, operation proceeds as in Figure 5, where all three visual modes accurately assess head position, and the inferred overall position serves to smooth noisy estimates.

In the top rows of Figure 6, at least one of the modes provides unreliable estimates: In Row 1, the subject faces away and loses color blob tracking; in Row 2, a jolt to the camera causes the background to become unreliable; in Row 3, a moving distractor weakens the foreground estimate and draws motion tracking away; and in Row 4, the lights are turned off, significantly degrading both the background- and color-based estimates. In all such cases, depressed reliability estimates in the corresponding modalities cause the system to weight other reports more heavily.

In Figure 7, we show modes of failure. Because hands and face masks exhibit properties similar to human heads (at least with respect to the three modalities), all three modes produce estimates that are deemed reliable. Additional computation to identify image regions as "head" or "not head" may alleviate this problem, although the case for masks and photographs is likely to require more subtle analysis.

Finally, we compare the estimates from Bayesian fusion against each of the components and a simple mean of the three component modality estimates. In Figure 8(top), we show the tracking estimates for the x positional estimate along with ground-truth data from the Polhemus tracker.

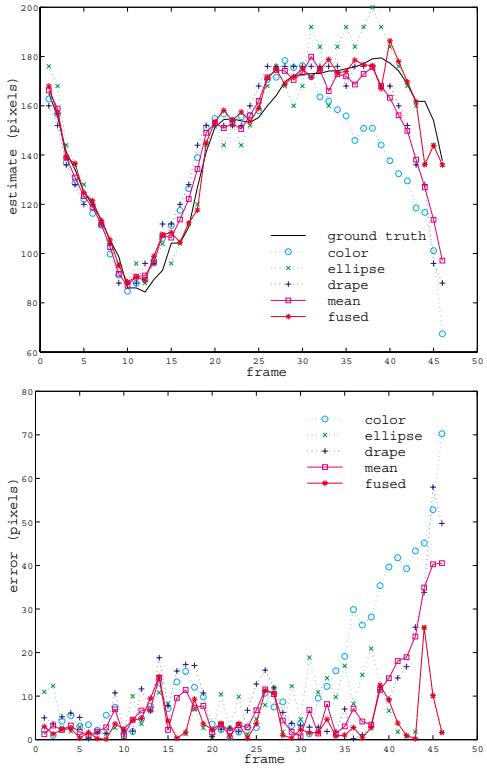

Figure 8. Comparison of Bayesian fusion with its constituent modalities, a simple mean, and ground truth.

Figure 8(bottom) shows the absolute error of each estimate when compared with ground truth. The particular image sequence included side-to-side head movement and for the last half of the sequence, sudden changes in illumination, effected by turning on and off a lamp.

We have found that Bayesian modality fusion usually outperforms any of its constituent modalities, often making estimates close to the modality estimate with the least error. Second, we see that even in moments where two out of the three modalities are widely off the mark (last several frames), the fusion algorithm remains close to its most reliable estimate. This last point suggests that Bayesian fusion is likely to outperform naive voting schemes. Certainly, it outperforms the simple mean, which is drawn away from ground truth by the two errant modalities. We can confidently say that Bayesian modality fusion performs an "intelligent" probabilistic weighting of the three modalities based on collected training data.

7 CONCLUSION

Bayesian modality fusion offers an expressive framework for weighting and integrating the reports from multiple visual modes. Our investigation has shown that Bayesian fusion of multiple modalities can generate reliable estimates of head position even in situations where component analyses are unreliable.

We foresee future work on the head-tracking system proceeding in several directions. First, we note that in order to

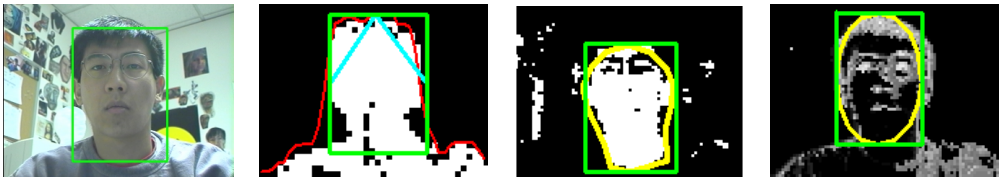

Figure 5. Successful head tracking under "normal" conditions. From left to right: the original color image with a rectangular overlay indicating the final state estimate, the background-subtracted image, the color-classified image, and the motion decay image.

fully train the system, training data proportional to all possible combinations of states is required. Even for our relatively simple network, we had to resort to default values in order to fill in the gaps. Situations for which training data was sparse inevitably caused more difficulty for fusion. Methods such as Gibbs' sampling allow for a more principled approach to estimating conditional probabilities even with sparse data, and application of such approaches is likely to enhance fusion, especially for events that occur seldomly.

We are also exploring the use of Bayesian network structure learning algorithms [5, 13] to learn the dependencies among key variables, rather than relying on learning solely to instantiate the parameters of a handcrafted dependency model. Learning algorithms have the ability to identify the *best* dependency model to use to infer location from evidence provided by multiple visual processing modalities. They can also characterize the strengths of the different dependencies, giving us advice about the relative value of modalities.

References

[1] Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, 1988.

[2] Y. Bar-Shalom and X.-R. Li. *Multitarget-Multisensor Tracking: Principles and Techniques*. YBS, 1995.

[3] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *Proc. Computer Vision and Patt. Recog.*, pages 232–237, 1998.

[4] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 33–42. AUAI, Morgan Kaufmann: San Francisco, July 1998.

[5] G. F. Cooper and E. Herskovits. A Bayesian method for constructing Bayesian belief networks from databases. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pages 86–94. Morgan Kaufmann: San Francisco, 1991.

[6] P. Dagum, A. Galper, E. Horvitz, and A. Seiver. Uncertain reasoning and forecasting. *Int'l J. of Forecasting*, 11(1):73–87, March 1995.

[7] P. Dagum, A. Galper, and E. Horvitz. Dynamic network models for forecasting. In *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, pages 41–48, Stanford, CA, July 1992. Association for Uncertainty in Artificial Intelligence, Morgan Kaufmann.

[8] T. Darrell, G. Gordon, J. Woodfill, and M. Harville. A virtual mirror interface using real-time robust face tracking. In *Proc. Int'l Conf. on Autom. Face and Gesture Recog.*, pages 616–621, 1998.

[9] J.W. Davis and A.F. Bobick. The representation and recognition of action using temporal templates. In *CVPR97*, pages 928–934, 1997.

[10] Microsoft Belief Networks: Tools for Bayesian Inference. http://research.microsoft.com/msbn.

[11] Association for Uncertainty in Artificial Intelligence. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 1990-1998.

[12] Z. Ghahramani. Learning dynamic Bayesian networks. In C. L. Giles and M. Gori, editors, *Adaptive Processing of Temporal Information*, pages 168–197. Springer-Verlag, 1998.

[13] D. Heckerman. A Bayesian approach to learning causal networks. Technical Report MSR-TR-95-04, Microsoft Research, 1995.

[14] E. Horvitz and M. Barry. Display of information for time-critical decision making. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 296–305, Montreal, Canada, August 1995. Morgan Kaufmann: San Francisco.

[15] E.J. Horvitz, J.S. Breese, and M. Henrion. Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning*, Special Issue on Uncertain Reasoning, 2:247–302, 1988.

[16] M. Isard and A. Blake. ICondensation: Unifying low-level and high-level tracking in a stochastic framework. In *Proc. European Conf. on Computer Vision*, pages I:893–908, 1998.

[17] F. V. Jensen. *An Introduction to Bayesian Networks*. Springer-Verlag, 1996.

[18] K. Kanazawa and T. Dean. A model for projection and action. In *Proceedings of the Eleventh IJCAI*. AAAI/International Joint Conferences on Artificial Intelligence, August 1989.

[19] L. Klein. *Sensor and Data Fusion Concepts and Applications*. SPIE, 1993.

[20] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[21] C. Rasmussen and G. Hager. Joint probabilistic techniques for trackin multi-part objects. In *Proc. Computer Vision and Patt. Recog.*, pages 16–21, 1998.

[22] H. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. In *Proc. Computer Vision and Patt. Recog.*, 1998.

[23] K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. *PAMI*, 20(1):39–51, January 1998.

[24] K. Toyama. Radial spanning for fast blob detection. In *Proc. Int'l Conf. on Comp. Vision, Patt. Recog., and Image Proc.*, 1998.

[25] M. Turk. Visual interaction with lifelike characters. In *Proc. Automatic Face and Gesture Recognition*, 1996.

[26] C.R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. Patt. Anal. and Mach. Intel.*, 19(7):780–785, 1997.

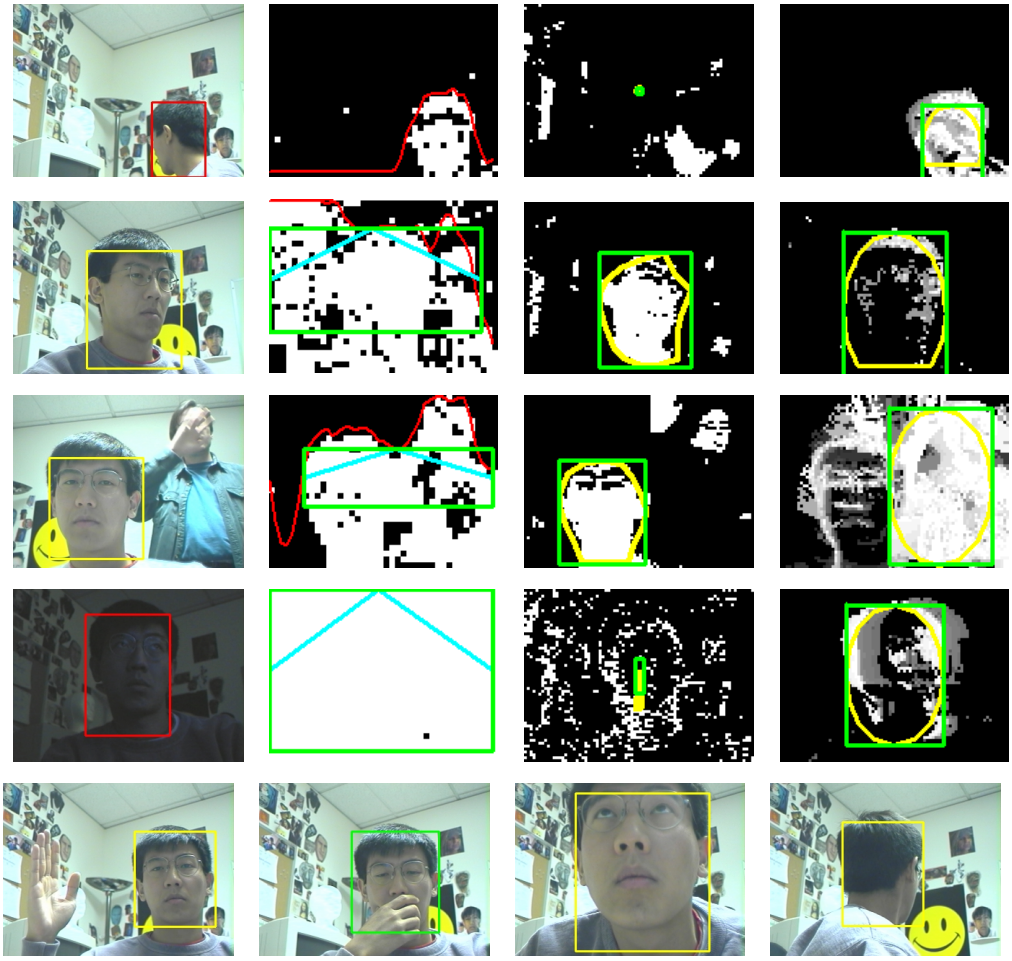

Figure 6. Successful head tracking under stress. The top four rows show instances where at least one of the modes fail (facing away; jolted camera; background distraction; lights out). The bottom row shows other cases of successful head tracking (raw image and final estimate only).

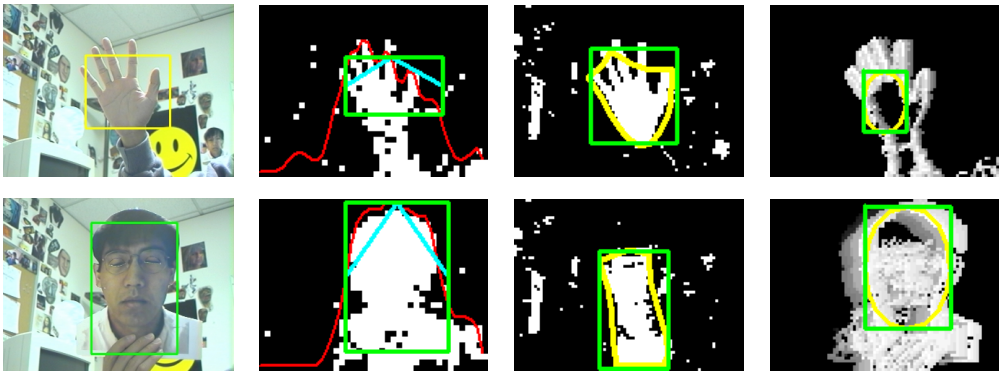

Figure 7. Failure modes in head tracking.