Sequential Monte Carlo Fusion of Sound and Vision for Speaker Tracking

J. Vermaak, M. Gangnet, A. Blake and P. Pérez
Microsoft Research Cambridge, Cambridge CB2 3NH, UK
Web: http://www.research.microsoft.com/vision

Abstract

Video telephony could be considerably enhanced by provision of a tracking system that allows freedom of movement to the speaker, while maintaining a well-framed image, for transmission over limited bandwidth. Already commercial multi-microphone systems exist which track speaker direction in order to reject background noise. Stereo sound and vision are complementary modalities in that sound is good for initialisation (where vision is expensive) whereas vision is good for localisation (where sound is less precise). Using generative probabilistic models and particle filtering, we show that stereo sound and vision can indeed be fused effectively, to make a system more capable than with either modality on its own.

1. Introduction

We establish design principles and demonstrate a working system that fuses stereophonic sound localisation with active contour tracking. Where more ambitious systems with several microphone pairs and several cameras, possibly steerable, could potentially handle free format multi-speaker interactions, here we aim at something more modest. A single, fixed camera with a single collocated microphone pair is well suited to video telephony, serving one or perhaps two speakers in a simple, closed environment. The setup of camera and microphones is illustrated in figure 1.

The processing of stereo sound is based on cross-correlation of the signal pair as a means of analysing Time Delay of Arrival (TDOA). In acoustic environments with relatively low noise and reverberation, triangulation based on the TDOA of measurements at a microphone pair [5, 12] is effective. In even moderately reverberant conditions, problems arise in that no unique TDOA can be determined. Some heuristic modifications to reduce the effects of reverberation have been proposed in *e.g.* [4, 6, 14], but these are reliant on either specific array configurations, or rather strong assumptions about the source signals and acoustic environment, and are far from robust in general scenarios. The alternative pursued here is to acknowledge that the TDOA D cannot uniquely be determined, to record a sequence $\{D_i\}$ of candidate TDOAs, and to model them jointly as probabilistic observations in clutter.

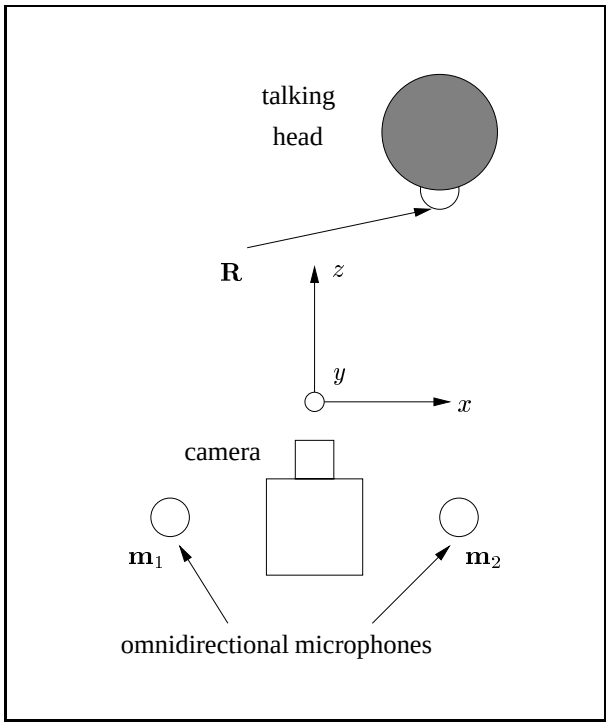


Figure 1. Audiovisual setup. A single microphone pair is positioned laterally and symmetrically with respect to the camera's optical axis, with its baseline in a horizontal plane.

The visual tracking uses a standard approach, based on a generative model for motion in a suitable contour state-space, together with a likelihood based on one-dimensional feature searches against a background of image clutter [8]. The probabilistic modelling of visual observations along a line is analogous to the processing of sound-signal cross-correlation peaks along the time-delay axis, in that both deal with linear search against a background of random clutter.

A particle filter is applied to fuse predictions from the generative model with aural and visual observations. This results in a tracking capability whose robustness is enhanced relative to vision alone. The sound information provides for initialisation, and helps considerably with recovery from loss of lock, as we demonstrate.

2. Observation Model for Sound

The sound measurement system consists of a pair of omnidirectional microphones as in figure 1, situated at positions $\mathbf{m}_1$ and $\mathbf{m}_2$ in the horizontal plane $y = 0$.

2.1. Time Delay of Arrival (TDOA)

The maximum TDOA that can be measured is $D_{\max} = c^{-1}\|\mathbf{m}_1 - \mathbf{m}_2\|$, with c the speed of sound (normally taken to be 342 ms^{-1}), and $\|\cdot\|$ the Euclidean norm. The true TDOA is given by

$$D = c^{-1}\left(\|\mathbf{R} - \mathbf{m}_1\| - \|\mathbf{R} - \mathbf{m}_2\|\right), \qquad (1)$$

where $\mathbf{R} = (x, y, z)$ is the source location. Apart from the true source, "ghost sources" due to reverberation lead to additional correspondences between left and right audio signals. These show up as additional peaks in the generalised cross-correlation function (GCCF) [9], as in figure 2.

Rather than trying to eliminate the spurious peaks, for example by further signal processing, they are acknowledged explicitly, knowing that the particle filter mechanism used for tracking and fusion is quite capable of assimilating them. The audio observation vector is therefore defined to be $\mathbf{z}_A = (D_1, \ldots, D_N)$, the N candidate TDOA measurements corresponding to the delay timings of the peaks of the GCCF. In what follows N is also considered unknown. Peaks not due to the true source are regarded as clutter.

2.2. Likelihood Model

A complete derivation and discussion of the likelihood model for the TDOA D can be found in [13]. The resulting likelihood follows from a multi-hypothesis analysis [2] under the assumption of mutual independence of the TDOA measurements, and is given by

$$L\left(\mathbf{z}_A \middle| D\right) \propto \frac{q_0}{2D_{\max}} + c\sum_{i=1}^{N} q_i \mathcal{N}\left(D_i; D, \sigma_D^2\right) \mathbb{I}_{\mathcal{D}}(D) \qquad (2)$$

if speech is present, and $L\left(\mathbf{z}_A \middle| D\right) \propto 1$ if it is absent, so that no influence is exerted by the audio stream in this case. In (2), q_0 is the prior probability of all measurements being due to clutter, q_i, $i = 1, \ldots, N$, is the prior probability of the i-th measurement corresponding to the true TDOA, c is a normalising constant, and $\mathbb{I}_{\mathcal{D}}(\cdot)$ is the indicator function for the set $\mathcal{D} = [-D_{\max}, D_{\max}]$. The variance σ_D^2 depends on the signal-to-noise ratio and the reverberation time of the acoustic environment, and can be set empirically. However, the performance of the tracking algorithm proved to be robust to the accuracy of the value chosen for this parameter.

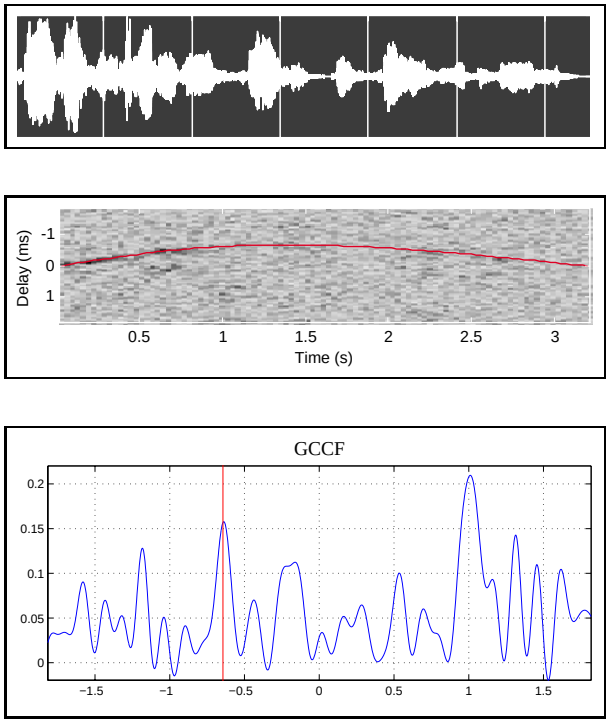


Figure 2. Reverberation generates multiple correspondences. A speech signal of around 3 seconds duration (top) gives a correlogram with multiple peaks (middle), shown by dark blobs. The true delay trajectory is shown overlay. A slice of the correlogram (bottom) at $t = 0.5s$, shows multiple peaks, and the peak of highest magnitude is not in fact the true peak (marked by a vertical line).

3. Observation Model for Vision

The observation model for vision is summarised here, although we do not go into full detail as the model follows largely established practice for active contour tracking.

3.1. Image Processing

Following standard practice in visual search algorithms (see *e.g.* [10]), visual measurements are taken along lines normal to an outline curve C, as in figure 3.

Point features along normals $\{\mathbf{n}^{(j)},\ j = 1, \ldots, M\}$ are defined by sampling the intensity function regularly along

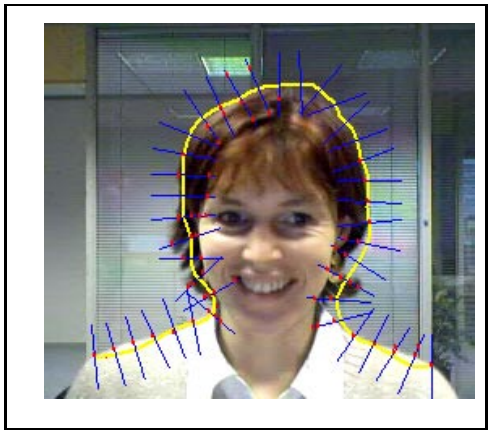


Figure 3. Image observations in clutter. Image measurements are made along lines normal to a hypothesised head contour C, as shown.

the line using bilinear interpolation of pixels, convolving with the gradient of a Gaussian (with width w approximately 1 pixel), and marking those maxima of response that exceeds a gradient threshold g. The resulting features have offsets $\nu^{(j)} = \{\nu_i^{(j)},\ i = 1, \ldots, N_j\}$, where an offset $\nu = 0$ indicates a feature lying on the hypothesised contour C. The combined image measurement is then $\mathbf{z}_I = \{\nu^{(j)},\ j = 1, \ldots, M\}$. It is important for inter-frame stability that g is defined relative to global image statistics, rather than local statistics gathered along one normal, or from the normals of one outline curve (which would be economical, computationally). Instead, independent samples of image gradient $|\nabla I|$ are taken distributed evenly over the image, and g is set to retain a proportion k_g of the strongest responses. Typically g is set to give $k_g = 30\%$. This imparts a measure of invariance to global illumination changes and drift in camera gain.

3.2. Image Likelihood

Likelihood modelling for observations along an individual normal $\mathbf{n}^{(j)}$ is straightforward, following similar reasoning as in the case of delay measurements for the sound signals above, to give a likelihood

$$L\left(\nu^{(j)} \middle| C\right) \propto q_0 + \frac{1-q_0}{N_j} \sum_{i=1}^{N_j} \mathcal{N}\left(\nu_i^{(j)}; 0, \sigma_I^2\right),$$

where q_0 is the non-detection probability for the visual contour (independent of q_0 for sound), and which is typically set to $q_0 = 0.2$ for reasonable behaviour. The variance σ_I^2 for valid contour measurements, assumed Gaussian, is determined from the residuals of contour fits on a few training images. Finally, the global image likelihood is computed as a product

$$L\left(\mathbf{z}_I | C\right) \propto \prod_{j=1}^{M} L\left(\nu^{(j)} \middle| C\right)$$

assuming joint independence of the $\nu^{(j)}$, and this is something for which some experimental justification has recently been claimed [11].

4. State-Space Model

In this study an image-based configuration space is used, on the camera-plane (x, y), rather than in 3D coordinates. This does not allow head rotation to be fully modelled, but that is outside the scope of a system that tracks only a bounding contour, as reported here, and the audiovisual calibration of an image-based system is more straightforward than the full 3D case.

4.1. Configuration Space

The image based configuration $X = (\mathbf{r}, T)$ consists of the image coordinates $\mathbf{r} = (x, y)$ of the centroid of a head-outline template, and the template itself as a curve $\mathbf{r}_0(s)$, obtained by drawing around a single-frame head, and which is perturbed affinely by T (a 2×2 matrix):

$$\mathbf{r}_X(s) = \mathbf{r} + (T + I)\,\mathbf{r}_0(s).$$

Further variability could easily be introduced using key-frames [3], but affine variability suffices for the experiments reported here. The head-outline template $\mathbf{r}_X$ is exactly the curve C used to obtain the visual measurements in section 3. The image based configuration used here does not allow the direct use of (1) to compute the TDOA D, since the 3D position $\mathbf{R}$ corresponding to a hypothesised configuration X is not uniquely determined. However, the geometry of the setup allows D to be computed from x using the Fraunhoffer approximation $D = D_{\max}\cos(\arctan(f/x))$, where f is the focal length of the camera, for which a pinhole model is adopted.

4.2. Dynamical Model

A stratified dynamical model was used to reflect the different kinds and degrees of variability that are appropriate to the head tracking task. The greatest variability is in horizontal motion (x-coordinate), followed by vertical motion (y), neither of which should be drawn towards any particular origin in the image, but which should remain within the field of view. Shape variability (T) is more constrained

— of smaller magnitude, and with a restoring tendency towards the home template $\mathbf{r}_0$. Dynamical models that reflect this are as follows, expressed discretely with respect to a sampling time interval τ (video frame-rate).

The displacement process is modelled as Langevin motion [1], as for a free particle in a liquid $\ddot{x}(t) + \beta^{(x)}\dot{x}(t) = w(t)$ with thermal excitation $w(t)$. The parameters of such a process are most naturally specified in terms of continuous-time parameters which have clear physical interpretations: the rate constant $\beta^{(x)}\, s^{-1}$, and the steady-state root-mean-square velocity $\bar{v}^{(x)}\, ms^{-1}$. It corresponds to a discrete process

$$u_t = a^{(x)} u_{t-1} + b^{(x)} w_t^{(x)} \quad x_t = x_{t-1} + \tau u_t$$

in which $w_t^{(x)}$ are $\mathcal{N}(0,1)$ variables and

$$a^{(x)} = \exp\left(-\beta^{(x)}\tau\right) \text{ and } b^{(x)} = \bar{v}^{(x)}\sqrt{1-(a^{(x)})^2}.$$

In experiments, we fixed $\beta^{(x)} = \beta^{(y)} = 10s^{-1}$ and $\bar{v}^{(x)}$ and $\bar{v}^{(y)}$ to 10% and 5% of the field of view in the respective directions, per second.

For the affine matrix, the model follows a stable, critically damped 2nd order autoregressive process, whose parameters are specified by a temporal rate constant $\beta^{(T)}$ and steady state root-mean-square magnitude $\bar{\rho}^{(T)}$ (dimensionless). It takes the discrete form

$$T_t = a_1^{(T)} T_{t-1} + a_2^{(T)} T_{t-2} + b^{(T)} w_t^{(T)},$$

in which $w_t^{(T)}$ are $\mathcal{N}(0,1)$ variables, and with $a_1^{(T)}$, $a_2^{(T)}$, $b^{(T)}$ set in terms of $\beta^{(T)}$, $\bar{\rho}^{(T)}$ according to well-known rules [3, p. 206]. We set $\beta^{(T)} = 10s^{-1}$ and $\bar{\rho}^{(T)}$ to 10%.

5. Particle Filter Tracking Algorithm

The general tracking problem involves the recursive estimation of the filtering distribution $p(X_k|\mathbf{z}_{1:k})$, with $\mathbf{z} = (\mathbf{z}_A, \mathbf{z}_I)$ and the subscript $1:k$ denoting all the observations from time 1 to time k, from which estimates of the configuration X can be obtained. The general recursions to compute the filtering distribution are given by

$$p\left(X_k|\,\mathbf{z}_{1:k-1}\right) = \int p\left(X_k|\,X_{k-1}\right) p\left(dX_{k-1}|\,\mathbf{z}_{1:k-1}\right)$$
$$p\left(X_k|\,\mathbf{z}_{1:k}\right) \propto L\left(\mathbf{z}_{A,k}|\,D_k\right) L\left(\mathbf{z}_{I,k}|\,C_k\right) p\left(X_k|\,\mathbf{z}_{1:k-1}\right),$$

where the first, or prediction, step uses the dynamical model and the filtering distribution at the previous time step to compute the one-step ahead prediction distribution, which then acts as the prior for the configuration in the second, or update, step where it is combined with the likelihood to obtain the filtering distribution.

Due to the non-linearity and multi-modality inherent in the problem, the recursions above are analytically intractable. Under these conditions sequential Monte Carlo, or particle filtering, methods [7, 8] provide an attractive solution strategy. The particular particle filter architecture adopted here deviates from the standard particle filter, and makes the best use of the properties of the model. Since the sound likelihood depends only on D, which in turn depends on the configuration X only via the image x coordinate, the sampling is separated in two stages by "partitioned sampling" [11]. In the first stages samples for the x coordinate are generated from a proposal distribution which is a mixture of the dynamics for x and the sound likelihood, viewed as a distribution in x. These samples are then properly reweighted with the sound likelihood, and resampled to populate x regions with high probability under the sound likelihood. In the second stage the remaining components y and T are proposed from their corresponding dynamics, reweighted with the image likelihood, and resampled. Since the broadest variations in X are due to x and y, separating x and y leads to a considerable improvement in sampling efficiency, palpable as a reduction in the number of particles needed per time step.

6. Results

We illustrate our system on two test sequences. The first sequence starts with a subject moving slowly from the centre of the image to the left, while all the time being quiet. The subject then moves rapidly to the right, where it pauses and speaks, and then progresses back to the centre. The second sequence involves two subjects (A and B), both appearing in the image at the same time. The subjects take turns to speak, while all the time moving their heads around to some degree. In what follows we will refer to the first sequence as the "motion" sequence, and to the second as the "ping-pong" sequence.

We performed two particle filtering experiments, using 20 particles in all cases, on each of the sequences. In the first experiment only the visual measurements were used to perform tracking, while visual and sound measurements were combined in the second experiment. The results of the first experiment on the "motion" sequence is summarised by the key frames in top of figure 4. The particle filter successfully tracks the subject during the period of normal motion to the left, but loses track during the rapid motion to the right. The particles latch on to prominent features in the background and never recover the subject again. However, in the second experiment, where the visual measurements are combined with the sound measurements, the particle filter is able to immediately reinitialise on the subject as soon

as it speaks, and the subsequent tracking is successful, as is illustrated by the key frames in the bottom of figure 4.

Similar results were obtained on the "ping-pong" sequence, and are summarised by the key frames in figure 5. In the case where only visual measurements are used the particles remain focussed on subject A, where they were initialised, regardless of which subject is speaking. When, on the other hand, the sound measurements are also used, the particles jump back and forth between the two subjects as they take turns in the conversation. Thus, the algorithm can be integrated into a teleconferencing system to determine the focus speaker for a steerable camera.

What is truly remarkable is that these results were achieved with low cost off-the-shelf equipment. The system was only very roughly calibrated, and proved to be robust to the exact values chosen for the intrinsic parameters of the camera, and did not require extremely careful placement of the microphones relative to the camera. Furthermore, no attempt was made to compensate for the reverberation and background noise, of which there was a fair amount due to fan and air-conditioner noise. Also, as is evident from the result sequences, tracking was performed against a cluttered background with many objects that can potentially distract a vision-only based tracking algorithm. Thus, it is proved that the combination of sound and vision can achieve a far more robust tracking performance, at a low computational cost, than any of the modalities on their own.

7. Conclusions

Further investigations are looking at the following issues.

- Ultimately a full 3D system may be desirable, so head rotation can be fully modelled, including displacement of the mouth relative to the centre of the head image.
- So far studies have been based on stored audiovisual sequences, but preliminary indications (based on software profiling) suggest that a real-time system should be quite feasible without special hardware, and work is currently in progress to achieve this.
- A more powerful system, on a teleconferencing scale, would use several microphones, or microphone pairs, distributed widely, not just at the camera centre This would require full three dimensional calibration which would be somewhat facilitated by the use of more than one camera also.
- It may be possible, and beneficial, to cut out the intermediate stage of marking correlation maxima for sound signal pairs, and evaluate a likelihood computed directly from the instantaneous value of the correlation (*i.e.* for one fixed delay D), if such a likelihood could satisfactorily be defined.

References

[1] K. Astrom. *Introduction to stochastic control theory*. Academic Press, 1970.

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

[3] A. Blake and M. Isard. *Active contours*. Springer, 1998.

[4] M. S. Brandstein. Time-delay estimation of reverberant speech exploiting harmonic structure. *Journal of the Acoustic Society of America*, 105(5):2914–2919, 1999.

[5] M. S. Brandstein and H. F. Silverman. A practical methodology for speech source localization with microphone arrays. *Computer, Speech and Language*, 11(2):91–126, 1997.

[6] M. S. Brandstein and H. F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 375–378, 1997.

[7] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F*, 140(2):107–113, 1993.

[8] M. Isard and A. Blake. Visual tracking by stochastic propagation of conditional density. In *Proc. 4th European Conf. Computer Vision*, pages 343–356, Cambridge, England, Apr 1996.

[9] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-24(4):320–327, 1976.

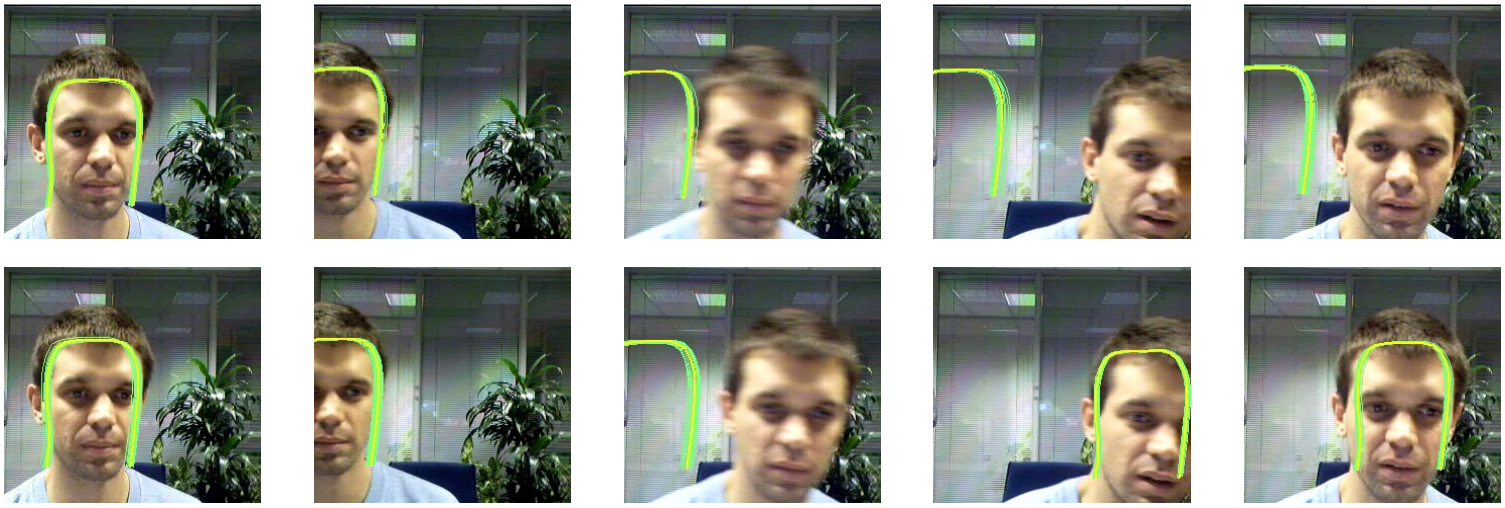
[10] D. Lowe. Robust model-based motion tracking through the integration of search and estimation. *Int. J. Computer Vision*, 8(2):113–122, 1992.

[11] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *Proc. Int. Conf. on Computer Vision*, pages 572–578, 1999.

[12] H. F. Silverman and E. Kirtman. A two-stage algorithm for determining talker location from linear microphone array data. *Computer Speech and Language*, 6:129–152, 1992.

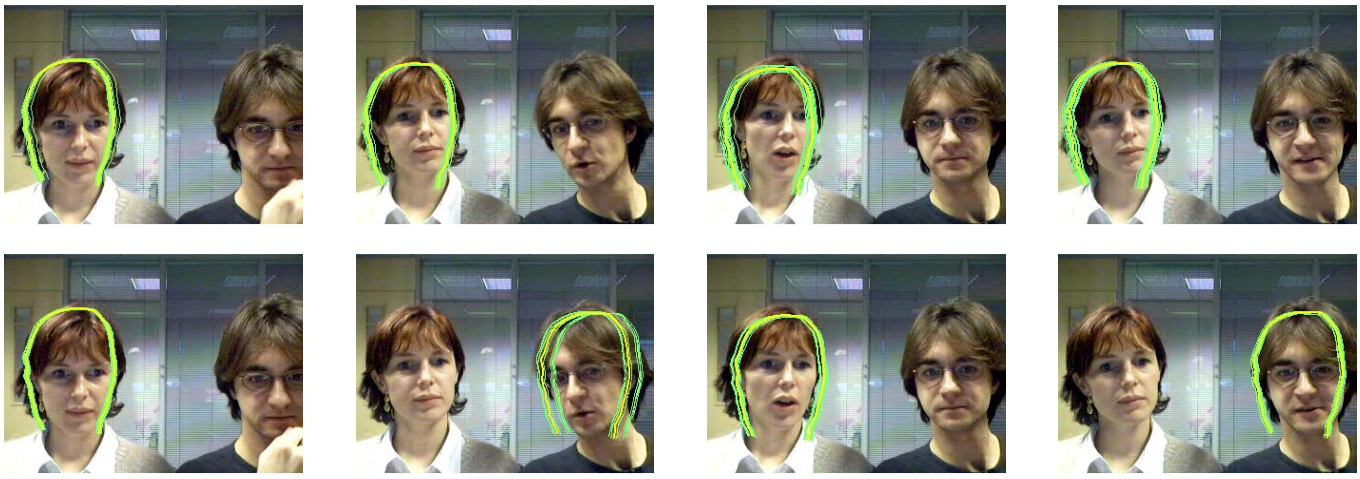
[13] J. Vermaak and A. Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, 2001.

[14] H. Wang and P. Chu. Voice source localization for automatic camera pointing system in videoconferencing. In *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 187–190, 1997.



0.24s 2.16s 3.40s 4.20s 6.12s

Figure 4. Recovery of lock. Subject moves rapidly (top), and using only vision to track, lock is lost, and does not recover. Using sound and vision (bottom), lock recovers.



0.64s 2.64s 4.32s 7.08s

Figure 5. Conversational ping-pong. Subjects speak alternatively. With vision only (top), tracking continues with initial subject. With vision and sound (bottom), tracking alternates with the speaker.