

Student- t Mixture Filter for Robust, Real-Time Visual Tracking

James Loxam and Tom Drummond

Department of Engineering, University of Cambridge

Abstract. Filtering is a key problem in modern information theory; from a series of noisy measurement, one would like to estimate the state of some system. A number of solutions exist in the literature, such as the Kalman filter or the various particle and hybrid filters, but each has its drawbacks.

In this paper, a filter is introduced based on a mixture of Student- t modes for all distributions, eliminating the need for arbitrary decisions when treating outliers and providing robust real-time operation in a true Bayesian manner.

1 Introduction

Filtering, estimating some hidden dynamical state over time from just noisy measurements of that state, in a robust, reliable manner is a challenging but important problem, with applications in many fields. Bayes' Theorem provides the mathematical framework for solving this problem, however design decisions impinge upon the implementation of this framework which put limitations on the performance of the final filter.

In this paper, an implementation is proposed based on a Student- t mixture model, which provides robust performance in the face of outliers at a speed allowing real-time operation on complex problems.

1.1 Background

Robust methods of estimation have been well covered in the literature, with most methods falling into the class of *M-estimators* [1]. M-estimation as a technique can be adapted by the use of different weighting functions (e.g. Huber, Cauchy and Tukey among others) to create likelihood functions with different properties, the maximum of which can then be found to provide a solution.

The problem with estimation in general, however, is that the computational load increases linearly with the number of measurements, and thus time. Filtering in a recursive manner (see section 2) circumvents this problem by maintaining a probability distribution over the state which encodes all previous measurements and can be simply updated at each timestep. Doing this in a robust manner, however, is non-trivial as the probability distributions must be proper (integrate to unity), thus precluding and method which uses a uniform component to model outliers (e.g. Tukey weighting function).

Filtering Methods The Kalman filter [2] is often considered the forerunner to modern filtering systems. In the case of normally distributed state and noise distributions, it is optimal. The real world, however, often produces noise which is far from normally distributed, e.g. measurement noise from a point tracker where data association sometimes fails or the process noise involved with flying a model helicopter in windy conditions, where occasional gusts of wind can cause rapid changes to the system state. In such systems, it is known that the Kalman filter performs poorly.

Due to the mathematical niceties of the normal distribution, many attempts have been made to improve performance, e.g. pre-filtering measurements using RANSAC [3] to remove any erroneous measurements. This often involves, however, rather arbitrary decisions about which measurements are erroneous and which are not and problems can still occur when these decisions turn out to be incorrect.

It has long been proposed that there should be no need to make such arbitrary decisions, that a properly formulated Bayesian approach should handle such problems [4]. The reason that the Kalman filter is unable to properly handle erroneous measurements is due to the light-weight tails it possesses, which effectively rule out the idea that any measurement is ever wrong.

Non-Parametric Filters The Bootstrap filter [5] and the Condensation algorithm [6] led the way in a new form of non-parametric filter. These non-parametric filters escaped the constraints of the normal distribution by representing the state as a set of discrete samples from that distribution, allowing the state to take arbitrary distributions and the use of any evaluable likelihood function (e.g. a Student- t distribution [7]). The problem with this early set of particle filters, however, was the computational cost, as the number of particles scaled exponentially with the dimensionality of the state.

Advancements in particle filters have mostly been due to improvements in the importance sampling densities used, such as in the Extended Kalman Particle Filter (EKPF) and the Unscented Particle Filter (UPF) [8], which provide a better distribution of particles, reducing the quantity required and thus the computational cost. More recent developments have thus turned into a class of hybrid filters or kernel-based particle filters, where parametric representations are used in parts of the algorithm to allow better particle placement [9, 10]. The restriction of using certain distributions to provide better particle placement, however, is the shift back towards limited distribution representation. By approximating all samples with a Gaussian distribution at each step (as in [9], or mixture of Gaussian distributions as in [10]) any semblance of heavy tails in the state distribution is being removed, eliminating the possibility of robustness to process noise outliers.

Although the approach in [11] is to use a Student- t kernel (and thus maintain the heavy tails), the system is limited to representing the posterior (and prior) by one Student- t , which fails to deal correctly with the problems associated with data confusion (see section 3.2).

Despite the improvements, particle and hybrid filters are still often expensive to run in high dimensional spaces and struggle to cope with both outliers in the process noise and measurement noise simultaneously.

Parametric Filters Many papers have discussed introducing heavy-tailed distributions into the filtering problem. Replacement of one of the noise distributions (either process noise or measurement noise) with a Student- t distribution has been introduced and has been shown to reject outliers in that noise distribution [12–14, 7]. Most early work, however, maintained a normal distribution for one part of the framework, and thus avoiding the problems associated with data confusion (see section 3.2).

The Gaussian Sum Filter (GSF) aims to approximate heavy-tailed distributions by using a mixture of normal distributions [15]. A theoretical problem with this is proved in [13] that for sufficiently large observations, the posterior is strictly unimodal (does not reject outliers) and thus the approximation breaks down. While this may be avoidable in a given application provided due care is taken, it is a limitation in using gaussian sums as a general method. Another practical limitation is the reduction of mixture size, complicated by the fact that different components of the mixture represent different parts of the distribution.

One of the first papers to discuss the use of Student- t distributions throughout the system was [16], albeit in a purely one-dimensional case. It discusses ideas such as *data confusion* (see section 3.2) but due to its methodology, is not extendable into higher dimensional space.

Section 2 shall discuss a basic framework for dynamical filtering followed by investigating the implications of using heavy-tailed distributions in section 3. Section 4 will introduce the base of the work followed by sections 5 and 6 which will describe the details of the work. Results are presented in section 7.

2 Filtering Theory

The filtering problem is concerned with determining, at a given time k , the posterior distribution of the state given all the previous gathered information, which comes in the form of measurements.

$$p(\mathbf{x}_k | \mathbf{Z}_k) \tag{1}$$

where \mathbf{x}_k represents the state distribution at time k and $\mathbf{Z}_k = \{\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_k\}$ represents the set of all measurements up to time k .

By formulating the filtering problem as a first-order Markov chain, the problem of finding the posterior distribution over all measurements can be reduced to the recursive update of a state estimate.

$$p(\mathbf{x}_k | \mathbf{Z}_k) = \frac{p(\mathbf{z}_k | \mathbf{x}_k, \mathbf{Z}_{k-1})p(\mathbf{x}_k | \mathbf{Z}_{k-1})}{p(\mathbf{z}_k | \mathbf{Z}_{k-1})} \tag{2}$$

$$p(\mathbf{x}_k | \mathbf{Z}_{k-1}) = \int p(\mathbf{x}_k | \mathbf{x}_{k-1})p(\mathbf{x}_{k-1} | \mathbf{Z}_{k-1})d\mathbf{x}_{k-1} \tag{3}$$

Formulating the problem recursively ensures the necessary property that the amount of computation required does not grow over time.

The different methods of filtering differ on their representation of these distributions; the Kalman filter represents all distributions by Normal distributions, making the equations above closed form, where as the particle filters tend to represent the prior and posterior distributions by sets of samples drawn from that distribution.

The main requirement when designing a filter is to ensure that it fits into this framework, and to ensure that, at each time step k , the posterior distribution has a constant structure to ensure recursive behaviour. When a non-conjugate prior is used then, approximations must be made to represent the posterior distribution.

3 Heavy-Tailed Distributions

For a long time, the Normal distribution has been used as the default distribution for approximating random variables. This has occurred due to its nice mathematical properties, it being closed under multiplication and convolution making it ideal for the Bayesian filtering framework mentioned above, and the ease with which it can be fit to data within an ML framework.

The Normal distribution is, however, also known for its lack of robustness to outliers, which comes as a consequence of its light-weight tails that die very quickly. Various stages of pre-filter (e.g. RANSAC) have been used to remove outliers ahead of the filtering, however with each of these forms there is a somewhat arbitrary decision about which measurements actually constitute ‘outliers’.

Heavy-tailed distributions have a non-negligible weight away from the mode, making the probability of outliers occurring non-negligible. By formulating the problem correctly and using appropriately heavy-tailed distributions, outliers need not be thought of as a special case, as they will simply be taken care of within the Bayesian framework.

3.1 Multivariate Student- t Distribution

The multivariate Student- t distribution represents a generalisation of the Gaussian distribution (the limit $\nu \rightarrow \infty$ produces the Gaussian distribution), and is defined as

$$S(x; P, \nu, \mu) = \underbrace{\frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})} \frac{|P|^{\frac{1}{2}}}{((\nu-2)\pi)^{\frac{d}{2}}}}_{1/z} \underbrace{\left(1 + \frac{\Delta^2}{\nu-2}\right)^{-\frac{\nu+d}{2}}}_{f(x; P, \nu, \mu)} \quad (4)$$

where $\Delta^2 = (x - \mu)^T P (x - \mu)$ is the squared Mahalanobis distance from x to μ and $\Gamma(\cdot)$ is the Gamma function. μ and P denote the mean and precision (inverse covariance) matrix respectively, while ν denotes the number of ‘degrees

of freedom” of the filter, which, in this instance, may take non-integer values. The dimension of the filter is given by d . The Student- t distribution allows the modelling of heavier tails, controllable by ν which can be directly related to Mardia’s measure of multivariate kurtosis [17].

$$\gamma_2 = \frac{2d(d+2)}{\nu-4} \quad (5)$$

The most common use of Mardia’s measure of kurtosis is to test the validity of Gaussian assumptions. Here, however, it will be used to parameterise the deviation of the Student- t modes from the Gaussian distribution.

3.2 Data Confusion

In Meinhold’s introduction of the Student- t distribution as the basis for in a filtering framework [16], he explicitly addressed the problem of *Data Confusion*, which had been ignored by earlier works. Data confusion is the tendency of heavy tailed distributions to generate multiple distinct modes under multiplication with each other (application of Bayes’ theorem) and is a natural consequence of heavy tailed distributions: logically by admitting that measurements can be ‘wrong’, a decision has to be made between them when they disagree. When both the likelihood and the prior have heavy-tails, as the two distributions diverge on their consensus of the current state, the posterior distribution becomes multimodal, effectively representing each of the two possibilities, one where the measurement was ‘correct’ and the prior ‘incorrect’, and the other vice-versa.

Although data confusion may initially appear to be a problem, it is merely reflecting reality and representing the current estimate as accurately as possible. Provided the distributions are dealt with as accurately as possible, future information will resolve the uncertainty when available, reinforcing one mode and not the other, which will then die away.

4 Student- t Mixture Filter

A mixture of Student- t distributions is proposed for the basis of a filter, both for the internal state estimate and the measurement noise distributions. This provides a heavy tailed distribution to deal with any outliers in a Bayesian manner, while allowing multiple modes to be propagated through time will enable it to deal with the problems attributed to data confusion. In order to maintain computation efficiency, only the N largest modes in the state estimate will be kept after each stage.

Since the Student- t distribution is closed neither under multiplication nor convolution, both the inference (equation 2) and the propagation of the state through time (equation 3) must be approximated to provide a recursive filtering framework.

5 Approximate Inference

Usual methods for approximating distributions involve the minimisation of some error metric (e.g. KL divergence or integrated square error), but due to the non-integrable form of the poly- t (product of Student- t distributions) distribution, none of these functions are analytically evaluable. Also, due to the multimodal nature of the posterior, moments do not carry sufficient information to make an approximation. As such, an alternate approximation scheme has been developed which concentrates on the key areas of the posterior while maintaining heavy tails away from the peaks.

Although it is impossible to give guarantees about the performance of this scheme, in practice performance is good. This is demonstrated by a number of experiments in section 7.1.

5.1 Mode approximation

Although the posterior poly- t distribution is neither integrable nor representable exactly by a sum of Student- t distributions, it does have a analytic representation, albeit with an unknown normalisation constant, C ,

$$p(\mathbf{x}|\mathbf{z}) = \frac{1}{C}p(\mathbf{x}) \prod_i^M p(\mathbf{z}_i|\mathbf{x}) \quad (6)$$

for M independent measurements, \mathbf{z}_i $i \in 1 \dots M$. As such, the value of the distribution and its derivatives can be calculated, up to scale. This is a key point to the method: a scaled approximation of the scaled posterior is calculated, which can then be normalised to provide a proper probability distribution (integrates to unity) and an approximation to the real posterior.

The proposed method of mode approximation is similar to that put forward in [10]. From a given set of starting points (see section 5.2), a Gauss-Newton optimisation is performed over the real (poly- t) posterior to find peaks in the state distribution. Gauss-Newton is particularly applicable in this situation as it does not require knowledge of the absolute scales of the Hessian and gradient of the cost function to operate, but merely their ratio, which can be calculated as it is independent of C . Once the locations of the peaks are known, a mode is placed at each.

$$\mu_i = \text{peak}_i \quad (7)$$

The *degrees-of-freedom* parameter can easily be determined by considering the decay rate of the real posterior. Each measurement has an exponent of $-\frac{1}{2}(\nu + d_z)$ (where d_z is the dimension of the measurement) which are summed when the product of the measurements is taken. For the multimodal prior, the decay rate is dominated by the smallest exponent, $-\frac{1}{2}(\min(\nu) + d_z)$. By equating this sum to the exponent of the approximating mode, a value for the *degrees-of-freedom* parameter of the approximating mode can be determined.

$$\nu_i = \min_{j \in [1, N]} (\nu_j) + \sum_{k=1}^M (\nu_k + d_z) \quad (8)$$

The precision matrix is approximated by the Hessian at the peak which this mode is representing. Since the actual Hessian of the posterior at this point cannot be calculated (due to the lack of a normalisation constant) the Hessians of the real posterior and the approximate mode are equated, each normalised by the value of their own distribution at this point. Once again, this is calculable for the poly- t posterior as it is a ratio, and allowing the precision of the approximate mode to be set.

$$P_i = -\frac{\nu - 2}{\nu + d} \frac{\nabla^2 p(\mathbf{x}|\mathbf{z})}{p(\mathbf{x}|\mathbf{z})} \Big|_{\mathbf{x}=\mu_i} \quad (9)$$

With all the parameters for the mode set, the weight of each mode can be set such that the peak of each approximate mode is the same height as the unnormalised poly- t posterior at that point.

$$w_i = \frac{p(\mathbf{x}|\mathbf{z})}{S(x; P_i, \nu_i, \mu_i)} \Big|_{\mathbf{x}=\mu_i} \quad (10)$$

Since $p(\mathbf{x}|\mathbf{z})$ was not a actual distribution and was only known upto scale, once all the modes have been estimated these weights then need to be normalised to provide a proper probability distribution for the posterior.

$$\sum_i w_i = 1 \quad (11)$$

5.2 Starting Points for Peak Finding

Since the posterior distribution is expected to contain multiple peaks, peak finding must be performed several times in an attempt to find them all. To maximise the chances of finding all of the peaks, the start points used for the search must cover as large a part of the space that is expected to contain peaks as possible.

A preliminary list of start points is generated from two sources.

- **Peaks in the Prior:** For each peak in the prior distribution, a search start point is generated. This set of start points should find all peaks corresponding to little change in the state, even in the presence of measurement outliers.
- **ML State estimates from random sets of measurements:** For a number of different randomly selected sets of measurements, an ML estimate of the state can be generated which can then be used as a start point. These start points are generated in much the same way as hypotheses in a RANSAC test. This set of start points should cover areas of the state space containing peaks due to correct sets of measurements, even when large changes to the state occur and the prior distribution is not close to the posterior.

The preliminary list of start points will often contain many overlapping points which would most likely converge to the same peaks. To avoid extra computation, and indeed to bound the amount of processing done, this preliminary list of start points is clustered using the *k-means++* algorithm [18]. From each cluster, the start point with the maximum probability (as evaluated in the real poly- t posterior) is used as a start point for peak finding for mode approximation.

5.3 Multimodal Measurement Distributions

During approximation of the posterior, the only properties requested of the real poly- t posterior (which is being approximating) is that the distribution can be evaluated and the first two derivatives calculated. As such, any measurement distribution could be used in the algorithm, provided these operations can be performed.

This provides the opportunity for using mixtures of Student- t s for the measurement distributions. Section 1.1 discussed how an explicit decision on outliers is not required in order to be able to deal with erroneous measurements; in a similar way by allowing multimodal measurement distributions to be incorporated the explicit one-to-one data association stage can be removed.

6 Approximate Time Propagation

Part of the initial Markov assumption was that of a model of how the state of the system will evolve over time and the application of transformation dynamics is the realisation of this model on the current state estimate.

For a known process model $g(\cdot)$,

$$x_{k+1} = g(x_k) \quad (12)$$

The first two moments have been well studied in the literature and the results of an application to a linear system are well known.

$$\mu_{x+1} = g(\mu_x) \quad (13)$$

$$\Sigma_{x+1} = (\nabla_x g) \Sigma_x (\nabla_x g)^T \quad (14)$$

In [17] it was also proved that Mardia's measure of kurtosis is invariant under non-singular transformations.

$$\gamma_{2:k+1} = \gamma_{2:k} \quad (15)$$

Use of an unscented transform instead of this extended transform, for increased accuracy in non-linear environments, is a trivial extension.

The state must then be convolved with the process noise distribution, to allow for errors in the model. The Student- t distribution is not closed under convolution and thus as with inference, the result must be approximated. Unlike multiplication, however, under convolution the Student- t distribution is guaranteed to be unimodal, which makes the matching of moments more meaningful.

Under the assumption of independence, it is well known that the first two standardised moments simply sum.

$$\begin{aligned} P(Y) &= P(X_1 + X_2 + \dots + X_N) \\ \mu_Y &= \sum_i \mu_i \quad \Sigma_Y = \sum_i \Sigma_i \end{aligned} \quad (16)$$

This falls out from the fact that these first two centralised moments are equal to the cumulants. Although the higher order cumulants still sum, since they do not equal the standardised centralised moments (e.g. skew, kurtosis), these moments do not sum. Using algebraic methods developed by [19] however, the following update equation has been derived.

$$\gamma_2(Y) = \frac{Tr(\Sigma_Y^{-2})}{d^2} \sum_i^N Tr(\Sigma_i^2) \gamma_2(X_i) \quad (17)$$

These equations can be used to perform simple moment matching to calculate the new state distribution for each mode in the state.

7 Results

7.1 Accuracy of Approximate Inference

Since the method of approximate inference introduced in section 5 is based on numerical methods, it does not provide a guaranteed level of performance. The results of tests indicating the actual level of performance are presented here. In each test, two Student- t distributions were generated with random parameters and multiplied together to produce a poly- t posterior. Two methods were then used to fit a mixture of two Student- t distributions to this poly- t posterior.

The first method used was a Maximum Likelihood (ML) estimator. 20000 samples were generated from the posterior distribution and a mixture of two Student- t distributions was fit to these samples using an iterative method designed to maximise the likelihood of the observed samples. The second method used was the method presented in section 5. The difference between these two approximations and the original poly- t posterior was then measured in terms of the KL divergence between the distributions.

The results for the ML fit indicate the suitability of using a mixture of Student- t distributions to approximate a poly- t distribution if one were allowed as much time as necessary to perform this approximation. The results for the method used in the Student- t Mixture Filter (SMF) illustrate the extra information lost in performing the approximation quickly using the method provided in section 5.

The results presented in table 1 show that in most cases, a mixture of Student- t distributions provide a good approximation to a poly- t distribution, as given by the low mean KL divergence. In addition, it also shows that the method presented causes a loss of very little extra information beyond that caused by the distributional approximation in the general case, as given by the low mean value. A shortfall of the proposed method is the maximum possible error, which occurs when two peaks interact with each other, changing the shape but not actually creating an extra peak. As shown by the results however, these occasions do not have an over-severe impact on the representation still having a relatively small KL divergence.

	Poly- t entropy	KL divergence from poly- t	
		ML	SMF
Mean	10.09	0.016	0.042
Standard Deviation	0.84	0.031	0.135
Maximum	-	0.109	0.756
Minimum	-	1.97×10^{-6}	2.15×10^{-6}

Table 1. KL divergence statistics from the actual poly- t distribution to a mixture of Student- t distributions fit by two different methods, a ML based parameter estimator and the method presented in section 5.

7.2 Visual Tracking

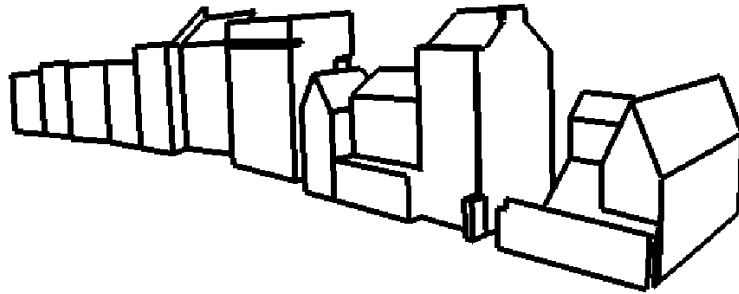


Fig. 1. Wireframe model of the street scene that was tracked.

Figure 1 shows a wireframe model of the street scene each of the filters was tasked with tracking. The model includes a number of point features (with known 3D location) and it is measurements of these that are supplied as input to each of the filters. For each frame, the FAST [20] feature detector is used to determine image features and normalised cross-correlation is used to match the features in the model to those in the current frame. The strongest matches are then supplied to the filters.

The Student- t filter is compared to both the Kalman filter and the EKPF across the sequences. The filters were set-up with constant position models. Measurement noise is assumed to be un-correlated, with 1-pixel variance in each direction. The EKPF was set-up to run with 100 particles.

Since the Kalman filter and the EKPF are known to be sensitive to outliers, RANSAC is used to remove outlying measurements prior to updating the filter to create a robust system. Note that this RANSAC stage is not needed with the Student- t mixture filter as it is inherently robust.

A video of tracking performance is supplied with the supplementary material as *outdoortracking.avi*, while figure 4 shows a number of frames from each of the trackers.



Fig. 2. A sample frame from one of the video sequences, with the model being tracked by the Student- t mixture filter.

Figure 3 presents the proportion of each sequence for which each filter was able to successfully track the buildings. It can easily be seen that the Student- t filter out-performs both the Kalman filter and the EKPF in terms of robustness: it maintains track for much longer than its competitors in each of the five sequences.

Filter	Measurement Update Time	Time inc. RANSAC
Student- t Mixture Filter	33.3ms	-
Kalman Filter	0.53ms	19.1ms
EKPF	40.3ms	58.8ms

Table 2. Average filter processing time per frame for each of the filters (Where RANSAC is included, 500 RANSAC tests were run). The times relating to the most robust implementation are shown in bold.

Table 2 shows the average update time required for the filter. As expected due to its simplicity the Kalman filter is the least computationally expensive, but at an average of $33.3ms$ per update the Student- t mixture filter falls within the bounds of frame rate operation. Additionally, the independent nature of the peak-finding process (the most expensive operation for the Student- t mixture filter) means the algorithm could easily be parallelised and run across multiple processors.

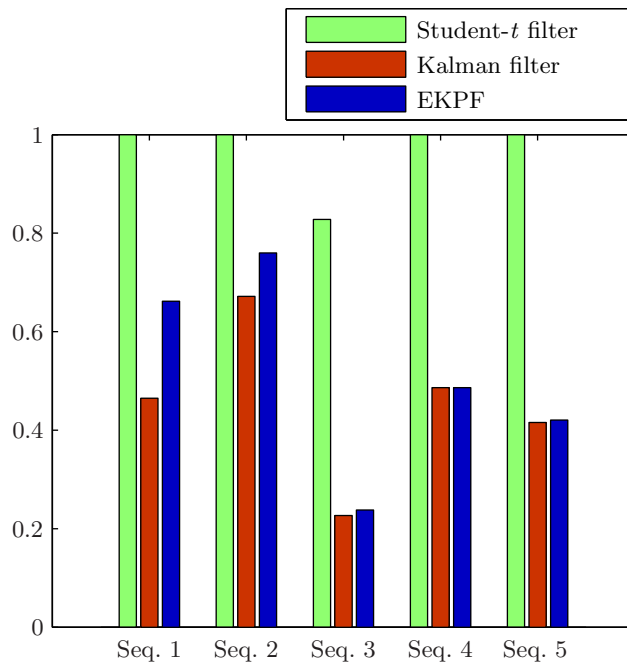


Fig. 3. Proportion of the sequence tracked by each filter for five different outdoor sequences. The Student- t filter can clearly be seen to out-perform both the Kalman filter and the EKPF in terms of robustness. Sequence 2 for each filter is included in the supplementary material for each filter, as is Sequence 4 for the Student- t mixture filter.

These timings are for each filter taking an input of up to 25 measurements. The Student- t Mixture filter simply took the top 25 matches by NCC score (of which 47% were outliers on average), whereas the Kalman filter and the EKPF had outliers removed beforehand, resulting in 14 inlying measurements per update on average. All filters scale with $O(m)$ for m measurements.

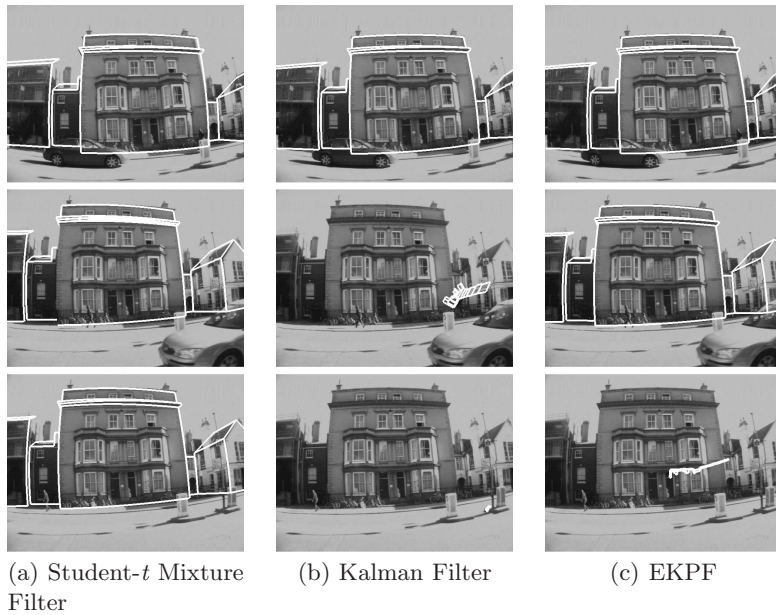


Fig. 4. Top to bottom, frames 100, 145 and 190 from Sequence 2 for each of the different filters tested. The Kalman filter and the EKPF lose track of the buildings at various points, whereas the Student- t mixture filter tracks the building for the entire sequence.

8 Conclusions

In this paper, the Student- t Mixture Filter has been introduced, a filter based around mixtures of Student- t distributions. The heavy tails of the Student- t provide an inherent robustness to ‘outliers’ in the measurement and process noise, negating the need for an explicit step to determine erroneous data. It has also been shown to outperform competing filters in real-world scenarios.

The Student- t Mixture Filter is also shown to be quick enough to run in real time, facilitating its use in real-time tracking systems. Being of a parametric base, it has polynomial complexity $O(n^3)$, thus presenting itself as a scalable solution for robust filtering.

References

1. Huber, P.: *Robust Statistics*. Wiley, New York (1981)
2. Kalman, R.: A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering* **82**(Series D) (1960) 35–45
3. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**(6) (1981) 381–395
4. Finetti, B.D.: The bayesian approach to the rejection of outliers. *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.* **1** (1961) 199–210
5. Gordon, N.: A hybrid bootstrap filter for target tracking in clutter. *IEEE Transactions on Aerospace and Electronic Systems* (Jan 1997)
6. Isard, M., Blake, A.: Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision* **29**(1) (1998) 5–28
7. Gordon, N., Smith, A.: Approximate non-gaussian bayesian estimation and modal consistency. *Journal of the Royal Statistical Society B* **55**(4) (1993) 913–918
8. van der Merwe, R., de Freitas, J., Doucet, A., Wan, E.: The unscented particle filter. In: *Advances in Neural Information Processing Systems 13*. (Nov 2001)
9. Kotecha, J., Djuric, P.: Gaussian sum particle filtering. *IEEE Transactions on Signal Processing* **51**(10) (Oct 2003) 2602–2612
10. Han, B., Zhu, Y., Comaniciu, D., Davis, L.: Kernel-based bayesian filtering for object tracking. *Proc. IEEE CVPR* (2005)
11. Li, S., Wang, H., Chai, T.: A t-distribution based particle filter for target tracking. *Proc. American Control Conference* (2006) 2191–2196
12. Dawid, A.P.: Posterior expectations for large observations. *Miscellanea* (1973) 664–667
13. O’Hagan, A.: On outlier rejection phenomena in bayes inference. *Journal of the Royal Statistical Society B* **41**(3) (1979) 358–367
14. West, M.: Robust sequential approximate bayesian estimation. *Journal of the Royal Statistical Society B* **43**(2) (1981) 157–166
15. Sorenson, H., Alspach, D.: Recursive Bayesian estimation using Gaussian sums. *Automatica* **7**(4) (July 1971) 465–479
16. Meinhold, R.J., Singpurwalla, N.D.: Robustification of kalman filter models. *Journal of the American Statistical Association* **84** (1989) 479–486
17. Mardia, K.: Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57** (1970) 519–530
18. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: *SODA ’07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Philadelphia, PA, USA, Society for Industrial and Applied Mathematics (2007) 1027–1035
19. Jammalamadaka, S.R., Rao, T.S., Terdik, G.: Higher order cumulants of random vectors, differential operators, and applications to statistical inference and time series. (1991)
20. Rosten, E., Drummond, T.: Machine learning for high speed corner detection. In: *9th European Conference on Computer Vision*. (May 2006)