Feature Tracking and Motion Classification Using A Switchable Model Kalman Filter

A. J. Lacey, N. A. Thacker and N. L. Seed
Dept. of EEE, University of Sheffield,
Mappin Street, S1 3JD, United Kingdom
email A.Lacey@sheffield.ac.uk

Abstract

In order to successfully track moving objects it is necessary to understand their motion. Such motion will inevitably change with time, thus attempting to fit the same model all of the time is inappropriate. An evaluation of the single model Kalman filter described in this paper demonstrates this. A maximum model coverage technique, however, would provide a intractable solution because of the exponentially increasing number of models required.

We present a solution which uses a finite set of models all making predictions on the data. One model is selected from the set on the basis that it best accounts for the data. A Kalman filter is then used to refine this model whilst it is appropriate. The remaining models continue to make predictions and at any time the filtered model may be replaced by one that more accurately describes the data. Experimental results demonstrate the improvement the switchable model Kalman filter provides over the single model Kalman filter. Problems using the chi-squared metric for model selection are discussed and a more appropriate metric, the Bhattacharyya integral, introduced. The desired properties of this integral are exposed via its analytic solution. It is believed that the statistical methods underlying this work may also have applications in other system identification problems.

1 Introduction

The goal of visual motion analysis is to develop a dynamic internal representation of the motion of objects in the world from an image sequence. In other words we wish to build a model of the observed motion and then use it to predict what the data will do in the future.

The problems that prohibit the recovery of accurate optical flow [1] and the constraints it places on the expected motion [2] mean it does not provide suitable data for model based tracking. It is our opinion that features present a more robust and less constrained set of object data for this type of tracking. The corner feature is particularly suited as it does not suffer from the aperture problem (as edges do).

BMVC 1994 doi:10.5244/C.8.59

The single model Kalman filter [3] has been used by many researchers to estimate model parameters based on the motion of visual features such as corners. If an accurate model of the system is supplied to the filter optimal model parameters can be achieved. However our experiments with the single model Kalman filter highlighted two common problems. First, when the motion of the feature changed dramatically but remained in accordance with the model, the prediction from the filter responded with an under/overshoot followed by exponential decay back to the correct prediction. The second problem was evident when the motion of the feature changed and could not be accounted for by the model. Such circumstances result is both inaccurate prediction and an unstable model.

Another common filter is the Lattice filter [4] and [5]. This is an alternative to the more common transversal filter and exploits a Toeplitz form of the autocorrelation matrix in order to solve the optimal filter normal equations. The assumed auto-regressive (AR) nature of the signal gives rise to an implementation where the filter weights or 'reflection coefficients' form a lattice structure between delayed and undelayed paths of the input signal. By monitoring the output of each lattice delay stage the order of the filter can be increased. However, over complex models can lead to unstable solutions and unnecessary prediction errors. Also the internal model generated is not a sensible model of the movement as individual parameters do not relate to physical motion characteristics [6]. A more practical problem with the filter is that it is necessary to provide data at discrete time intervals. This can cause problems when using image features because their detection is dependent on factors such as the detector performance and occlusion.

In the case where the origin of the data is uncertain, data combination filters such as the Probalistic Data Association Filter (PDAF) of Bar-Shalom [7] are commonly used. With such techniques, estimations generated from either a set of candidate data or a set of candidate models are combined as a weighted sum. The weights being the probabilities that each datum or model was correct. The resultant estimate is therefore based on total evidence. Estimations from the PDAF are generated using embedded single model filters typically the Kalman filter, therefore the problems discussed above can to some extent be attributed to the PDAF. Further the incoming data arose from only one source and therefore mixing predictions from multiple sources will introduce error.

What is needed is a filter capable of selecting the most appropriate model, based solely on the data, from a set of models. Selection of one model will not only reduce the error but will classify the motion of a feature. Further the model selection criteria should prefer the simplest model capable of predicting the data. This would ensure that the filter is representing the data in the most appropriate form, using the most stable model that predicts as accurately as the measurement system allows. We have therefore developed a switchable model filter based around the Kalman filter.

2 Development of Kalman filter algorithms

2.1 Single model Kalman filter

Our initial work involved a single model Kalman filter with a constant velocity model of motion. The representation of constant velocity used by the filter was as shown in appendix A. This representation is known as moving origin where the previous position prediction as well as current velocity estimation are available as state variables. The initial estimates of the state variables were generated using the constant velocity equations of appendix A, together with two frames of data from a single extracted corner feature. This initial state prediction was augmented with an initial estimate of covariance using the error propagation technique summarised in appendix B. This technique propagates the measurement error through the model providing an accurate initial estimate of covariance as opposed to an arbitrary 'big' one. This is something which is utilised in the switching algorithms discussed later. An artificial sequence of a single corner moving with stages of constant velocity and constant position was used to test the filter.

In order for the prediction to obey the underlying motion of the data and not attempt to predict the sporadic random noise overlaid upon it, the filter has a finite response time. This response time is automatically set by the filter using the specified error covariance and perturbation. Provided the incoming data obey the embodied model the predictions made by the filter are optimal in the mean squared error sense. However if, as is often the case in dynamic systems, after a period of time the observed data is drawn from a model other than the one described to the filter, the predictions will become incorrect.

The single model Kalman filter will always modify its parameters on the assumption that the data is consistent with the basis model it was using. We have found that this presents two problems. Firstly, if after a change in the motion parameters the data was still consistent with the embodied model a finite period of error prediction was incurred while the model parameters were modified. This 'sluggish' behaviour is a result of the finite response time. In order to reduce the time constant of this response it is common practice to modify the error covariance and/or perturbation until critical damping is achieved. However the response time is set by the filter in order to take account of the random additive noise. Therefore modifying it for a different system behaviour effects the filters ability to perform accurate data combination. Secondly, if the underlying motion changed completely so that the data was no longer consistent with the filter basis model the filter was unable to find a stable parameter set. This results in both large prediction error and model instability because fundamentally the model was wrong.

2.2 Switchable model Kalman filter

If we consider the problems of the single model filter we find they are both attributable to the assumption that all the data is drawn from the same model and that the parameters of that model change slowly relative to the noise variation. In order to overcome this assumption we suggest the addition of a parallel set of alternative models. Each model makes a prediction using the minimum amount of recent data. All predictions including the Kalman filter one are compared to the

next measurement. The model which most accurately predicts this measurement is selected for combination with subsequent data using the Kalman filter. This model is copied to the filter in place of the previous one. The rest of the models including the original of the winning model are reset ready to make predictions, together with the Kalman filter on the next frame of data. See figure 1.

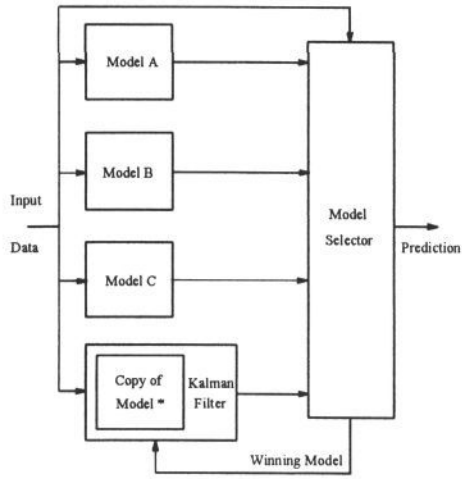

Figure 1: Switchable Model Kalman Filter Algorithm

The Kalman filter uses the model which best described the previous measurement. It refines the parameters of this model in accordance with the specified system statistics. If the data changes abruptly a more appropriate model can be switched in. This new model may be completely different from the current model if the data is better described that way. Or it may be the same basis model but with different parameters. This is possible because the model currently used by the Kalman filter is also available in unfiltered form. The unfiltered form is using the minimum amount of data needed to make a prediction and thus can respond rapidly to parameter changes. In this way we are more likely to have the most appropriate model for the data and therefore should make accurate predictions.

The problem remaining is that of how to make the model selection. The naive solution to this would be to select the model whose prediction is closest to the measurement. However, more complex models, having more free parameters, are able to predict a greater variety of data sets. This means that a complex model cannot predict any one data set as strongly as a simpler model. In other words we need to consider the complexity of the model, preferring the simplest model capable of accurately predicting the data, in order to ensure we have the most probable hypotheses for the data. Because model complexity is proportional to the width of the distribution this biased selection can be considered as an algorithmic embodiment of Occams razor.

The chi-squared, eqn 1, is commonly used to help fit a given model to a set of data and could be considered an obvious candidate for model selection as a test which confirms that the current data point is consistent with a particular model prediction. The form of the chi-squared shown includes a normalisation term.

This term can be ignored when using the chi-squared to fit model parameters to data as it is constant. However, in this case it is included because the model itself is changing.

$$\chi^2 = \frac{1}{\sqrt{2\pi(\sigma_m^2 + \sigma_p^2)}} \exp - \left[\frac{(\mu_m - \mu_p)^2}{2(\sigma_m^2 + \sigma_p^2)}\right] \quad (1)$$

Computation of the chi-squared requires measurement and measurement error. For any feature detection process the expected measurement error can be deduced or measured by experiment. We also need to compute a variance on the model prediction. For the unfiltered model predictions this is achieved by propagating the measurement error through the model as shown in appendix B. The Kalman filter already provides a covariance prediction for the state variables and so no extra computation is required. When compared to the measurement, selecting the model with the largest chi-squared could be considered as selecting the most appropriate description of the data. Or is it?

3 Bhattacharyya measure

The question that we should be asking when selecting a model is; which model will give the best prediction of the next data point based on the latest data point? We will assume that the model which provided the best constraint on the current data point will be most likely to give the best constraint on the next. The prediction power of a Kalman filter has as much to do with the width of the distribution of the prediction of the filter as the central value. Following this line of reasoning we can conclude that the best prediction that the Kalman filter could make for the distribution of the next data point would be the one which matched the distribution of the observed data, not just a close estimate in the sense of a chi-squared but a correctly bounded estimate. Given this interpretation we suggest the use of an alternative similarity metric, the Bhattacharyya integral [8], eqn 2. This integral has been used by other researchers in the field of pattern recognition as a two class separability metric [9]. It has also been used as a similarity metric [10] to verify the similarity of two distributions and it is in this context that we use it, see appendix C. However our interpretation of the result is not as an upper bound on the Bayes error as in [10], but as an absolute measure of distribution similarity as in [11].

$$-ln \int_{-\infty}^{\infty} \sqrt{PDF_m}\sqrt{PDF_p} \quad (2)$$

In order to best appreciate the Bhattacharyya metric the analytical solution of the integral for one dimensional Gaussian distributions is presented, eqn 3. The key steps of its derivation are given in appendix D.

$$\frac{\sqrt{2\sigma_m\sqrt{\sigma_m^2 + \sigma_p^2}}}{\sqrt{2\sigma_m^2 + \sigma_p^2}} \exp - \left[\frac{(\mu_m - \mu_p)^2}{4(2\sigma_m^2 + \sigma_p^2)}\right] \quad (3)$$

The form of the Bhattacharyya integral is similar to the chi-squared and like the chi-squared it weights the result in favour of the simplest model. However, the Bhattacharyya integral differs from the chi-squared in the normalisation term such that the integral is dimensionless. It is interesting to note that only when the square root of the probabilities are taken is the result of the probability overlap integral dimensionless, as it must be for a probablistic comparison measure.

For the case in two dimensions it can be shown that the solution is the addition of two one dimensional solutions, after taking their natural logarithm, (eqn 4).

$$\ln\left[\frac{2\sqrt{\sigma_{xm}\sigma_{xpm}\sigma_{ym}\sigma_{ypm}}}{\sqrt{\sigma_{xm}^2+\sigma_{xpm}^2}\sqrt{\sigma_{ym}^2+\sigma_{ypm}^2}}\right] - \frac{1}{4}\left[\frac{(\mu_{xm}-\mu_{xp})^2}{(\sigma_{xm}^2+\sigma_{xpm}^2)} + \frac{(\mu_{ym}-\mu_{yp})^2}{(\sigma_{ym}^2+\sigma_{ypm}^2)}\right] \quad (4)$$

$$\sigma_{xpm} = \sqrt{\sigma_{xp}^2+\sigma_{xm}^2} \quad , \quad \sigma_{ypm} = \sqrt{\sigma_{yp}^2+\sigma_{ym}^2}$$

This solution can only be used when the two distributions are uncorrelated. If they are not we are left with the integration of a cross term which has no direct analytic solution.

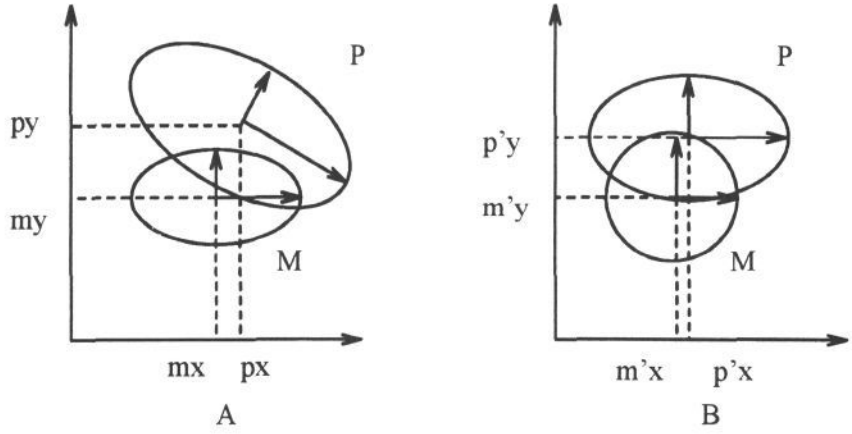

Figure 2: Transformation to Ensure Uncorrelated Errors.

3.1 Orthogonalisation

In order to ensure that this cross term does not exist we need to transform the data into a co-ordinate system where the two distributions are aligned with the parameter axes. The algorithm to perform this transformation is a variation on simultaneous diagonalization [12] and is as follows;

- Select a co-ordinate system, x and y, parallel to the measurement errors. These will be orthogonal co-ordinates by virtue of the measurement space.

- Scale the errors so that the x and y measurement error ellipse becomes a circle (not necessary when we are dealing with square pixels).
- Find the principle axes of the prediction error ellipse. These are the eigen vectors of the prediction covariance matrix.
- Using the principle axes to define the new co-ordinate system transform the data into this system.

The diagram of figure 2 demonstrates this transformation showing the orientations of the error ellipse before transformation, A, and after, B transformation. In this way the analytical form of the integral can be used to compare each prediction with the measurement, selecting the model with the largest similarity score.

4 Experiments and results

To compare the switchable filter to the single model filter and to compare the performance of the chi-squared and Bhattacharyya metrics a series of experiments were performed. For the switchable methods a second model, constant position, was added. Although simple the constant position model together with the constant velocity model provides sufficient information to form a complete description of the motion of the feature in the artificial sequence. Two forms of error were introduced to the recovered data. First a random position error from a Gaussian distribution of 0.3 pixel variance. This simulated inaccuracies in the detection process and is typical of corner detection [13]. Second, for every experiment the amount of recovered data supplied to the filter was varied from 100% down to 50%. This was done to simulate detection reliability, which for corner detection and correspondence solving is in the range 60%-80% [13] and visual occlusion which is entirely scene dependent. For each algorithm the experiment was repeated 100 times for every percentage of missing data. After each experiment the average prediction error per frame was calculated. This value does not include outlier predictions, classed as any prediction outside of 3 standard deviations from the measurement. The value displayed on the graphs is the average, over 100 experiments, of the average prediction error per frame. Its value is, therefore less dependent upon the location of the missing data.

The first graph of figure 3 shows the results from the single model and chi-squared switching algorithms. The improvement in prediction accuracy of the switching algorithm is obvious from the graph. The chi-squared performed 23.44 model switches with 100% of the data leading to 4.46 outlier points being identified. The single model method, believing all data is consistent with the model has a much larger expected error and thus identifies very few outliers only 0.02 on average with all the data. The second graph of figure 3 compares the results using the chi-squared and Bhattacharyya as switching metrics. The Bhattacharyya provides visible improvement over the chi-squared, particularly as the amount of missing data increases. The Bhattacharyya performed 19.52 model switches with 100% of the data leading to 3.91 outlier points being detected. The number of actual motion changes in the sequence was 5.00.

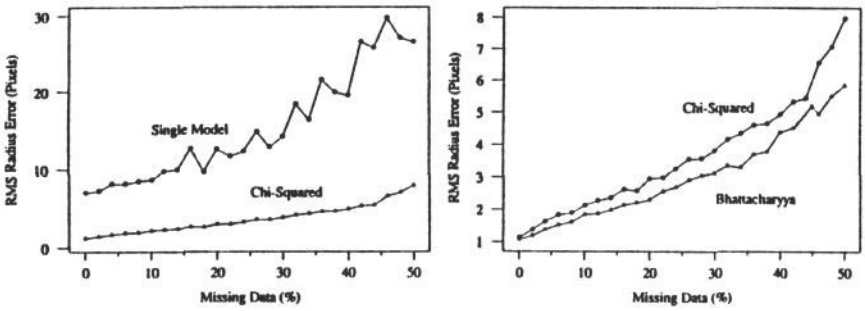

Figure 3: Prediction Error Graphs for the 3 Methods

5 Discussion and conclusion

The results show that switching can give superior results given that model coverage is sufficient. After a change in the observed motion a latency of a single frame is incurred, where the incorrect model is used to make a prediction, before the correct model is switched in. We have demonstrated the technique using only two simple models. However any number of models, each of differing complexity, can be used. Best results being obtained when the model set provides complete coverage over the expected motion.

In this paper we have discussed the need for on-line model selection when tracking moving features. To this end we have presented the finite memory switchable model Kalman filter. We have shown the need to prefer the simplest model which accurately predicts the data. Although the chi-squared metric embodies this property we believe that a better selection criterion can be based on trying to select the model which will give the most accurate constraint on the next measurement (i.e. the model which has the best prediction ability). We believe that the effectiveness of this constraint is embodied in the Bhattacharyya overlap integral, although this is not the conventional interpretation of this measure.

Experimental results demonstrate the improvement gained using the switchable model Kalman filter over the usual single model Kalman filter as well as the greater prediction accuracy achieved using the Bhattacharyya metric for model selection rather than the chi-squared. The ability of the switchable model Kalman filter not only to give improved prediction accuracy but also to classify the data in terms of one particular model leads us to believe that extensions of this statistical technique would have applications in other classes of system identification problem.

Appendix A. Constant velocity model

$$V_x = \frac{x(n+1) - x(n)}{\delta t} \qquad V_y = \frac{y(n+1) - y(n)}{\delta t}$$

$$x(n+1) = x(n) + V_x\delta t \qquad y(n+1) = y(n) + V_y\delta t$$

which can be formulated as $p(n+1) = Hp(n)$ where

$$H = \begin{vmatrix} 1 & 0 & \delta t & 0 \\ 0 & 1 & 0 & \delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix} \qquad p(n) = (x(n), y(n), V_x(n), V_y(n))^T$$

Appendix B. Error propagation

The estimate on the covariance on the initial estimate of the state variables C_p is given by

$$C_p = F^T C F$$

$$F = \nabla_x p(n+1) \ , \ X = (x(n),\ y(n),\ x(n+1),\ y(n+1))$$

where for $\delta t = 1$

$$F = \begin{vmatrix} 2 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ -1 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 \end{vmatrix} \qquad C = \begin{vmatrix} \sigma^2_{xn} & 0 & 0 & 0 \\ 0 & \sigma^2_{yn} & 0 & 0 \\ 0 & 0 & \sigma^2_{xn-1} & 0 \\ 0 & 0 & 0 & \sigma^2_{yn-1} \end{vmatrix}$$

Appendix C. Relationship to frequency coded distributions.

The standard maximum likelihood statistic for comparing two frequency distributions S_a and S_b is given by

$$\chi^2 = \sum_i^n (S_{ia} - S_{ib})^2/(S_{ia} + S_{ib})$$

Each frequency measure will be distributed as a Poisson which in the limit of large numbers will approximate Gaussian distributions with variance S_i.

For small variances the chi-squared statistic can be approximated in any alternative space for which there is a smooth continuous mapping function $F(S)$.

$$\chi^2 = \sum_i^n \frac{(f(S_{ia}) - f(S_{ib}))^2}{(\frac{\partial f(S_{ia})}{\partial S_{ia}})^2 S_{ia} + (\frac{\partial f(S_{ib})}{\partial S_{ib}})^2 S_{ib}}$$

In the special case of $f(S) = \sqrt{S}$ we obtain the Matusta distance measure.

$$\chi^2 = 4\sum_i^n (\sqrt{S_{ia}} - \sqrt{S_{ib}})^2$$

Mapped to this space all probability values have the same variance, the space has thus been linearised. Notice also that the possible infinities in the original definition are now finite.

We argue that in this space large scale differences between the two distributions are more meaningful than the original χ^2 definition. In the limit of probability distributions

$$\chi^2 = const - 8\sum_{i}^{n}\sqrt{P_{ia}}\sqrt{P_{ib}}$$

giving the Bhattacharyya distance measure as the similarity function. The conventional use of this measure is as an upper bound on the Bayes error for a two class assignment problem. Instead, we suggest the use of this measure not as a estimate of separability but as an absolute measure of similarity.

Appendix D. Analytical 1D Bhattacharyya

With Gaussian distributions the Bhattacharyya integral becomes

$$B_l = -ln\frac{1}{\sqrt{2\pi\sigma_a\sigma_b}}\int_{\infty}^{\infty} exp - \frac{1}{4}((x-\mu_a)^2/\sigma_a^2 + (x-\mu_b)^2/\sigma_b^2)dx$$

$$= -ln\frac{exp\frac{(\mu_a-\mu_b)^2}{4(\sigma_a^2+\sigma_b^2)}}{\sqrt{2\pi\sigma_a\sigma_b}}\int_{\infty}^{\infty} exp - \left(\frac{\sigma_a^2+\sigma_b^2}{4\sigma_a^2\sigma_b^2}\left(x - \frac{\sigma_b^2\mu_a + \sigma_a^2\mu_b}{\sigma_a^2+\sigma_b^2}\right)^2\right)dx$$

$$= -ln(\frac{\sqrt{2\sigma_a\sigma_b}}{\sqrt{\sigma_a^2+\sigma_b^2}}) + \frac{(\mu_a-\mu_b)^2}{4(\sigma_b^2+\sigma_a^2)}$$

References

[1] Verri, A. and Poggio, T., "Against Quantitative Optical Flow", 1st Int. Conference of Computer Vision, 1987.

[2] Aggarwal, J. K. and Nanrhakumar, N., "On the Computation of Motion from Sequences of Images: A Review", IEEE Proceedings, V. 76, No. 8, 1988.

[3] Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems", Journal of Basic Engineering, March 1960.

[4] Goodwin, G. C. and Sin, K. S., "Adaptive Filtering, Prediction and Control", Prentice-Hall, 1984.

[5] Alexander, S. T., "Adaptive Signal Processing: Theory and Applications", Springer-Verlang, 1986.

[6] Mayhew, J. E. W., Zheng, Y. and Billings, S. A., "Layered Architecture for the Control of Micro Saccadic Tracking of a Stereo Camera Head", BMVC, 1992.

[7] Bar-Shalom, Y. and Fortmann, T. E., "Tracking and Data Association", Mathematics in Science and Engineering, Volume 179, Academic Press, 1988.

[8] Bhattacharyya, A. "On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions", Bull, Calcutta Math. Soc. 35, pp 99-110, 1943.

[9] Devijver, P. R. and Kittler, J., "Pattern Recognition: A Statistical Approach", Prentice Hall, 1982.

[10] Tugnait, J. K. and Haddad, A. H., "A Detection-Estimation Scheme for State Estimation in Switching Environments", Automatica, V. 15, 477-481, 1979.

[11] Evans, A. C., Thacker, N. A. and Mayhew, J. E. W., "The Use of Geometric Histograms for Model Based Object Recognition." BMVC, 1993.

[12] Fukunaga, K. "Introduction to Statistical Pattern Recognition", 2nd ed., Academic Press, 1990.

[13] Thacker, N. A. and Courtney, P., "Statistical Analysis of a Stereo Matching Algorithm", BMVC, 1992.