GLOBAL MOTION ESTIMATION USING BLOCK MATCHING WITH UNCERTAINTY ANALYSIS

Pekka Sangi, Jari Hannuksela, Janne Heikkilä

Machine Vision Group, Department of Electrical and Information Engineering
P.O.Box 4500, FIN-90014 University of Oulu, Finland
phone: + (358) 8 553 1011, fax: + (358) 8 553 2612, email: {psangi,jhannuks,jth}@ee.oulu.fi
web: http://www.ee.oulu.fi/mvg

ABSTRACT

The paper presents an approach to dominant global motion estimation which is based on estimation of displacements of a small set of blocks. For each block, a robust motion-compensated block matching measure is evaluated for a fixed range of displacements. Analysis of matching measure values provides displacement estimates, and the related uncertainty is also evaluated using a gradient based thresholding method. The results of the uncertainty analysis are then utilized in global motion estimation for outlier analysis and parametric motion model fitting. The performance of the technique is evaluated in experiments which show the usefulness of the approach.

1. INTRODUCTION

Global motion estimation is an important subtask in many applications such as video coding, video stabilization and vision-based human-computer interaction. The goal is to determine a parametric model, which approximates well the dominating motion apparent between two images. The choice of model is dictated by the characteristics of expected motion and accuracy requirements. One common approximation is to use an affine model where local displacement is a linear function of image coordinates. Estimation of such a model can be based on direct use of image intensity information, which is done in gradient based methods [1], or one may use an indirect feature based method [2].

Motion estimation between images is typically an uncertain process due to several reasons. Changes in illumination conditions and image capturing parameters can cause corresponding regions to have different appearances. In some image region, similar texture may exist in several locations leading to ambiguity. In the case of global motion estimation, there might be other independent motions in the scene which should not affect the result. Thus, robust approaches are needed so that changes in imaging conditions can be tolerated, and misleading observations can be discarded. Moreover, in the embedded applications that we are interested in, the requirements on computational resources should be limited.

In this paper, we consider a feature-based approach to global motion estimation, which addresses these issues. The method can be viewed as a four-stage process (Fig. 1). In order to restrict the processing requirements, a sparse set of blocks is selected from the anchor image in the first step. Next, for each block, a block matching measure is evaluated for a certain range of displacement candidates. The results are analyzed to obtain block displacement estimates and the associated uncertainty information. In the last step, global motion is estimated by fitting a parametric motion model to the displacements selected in the outlier analysis.

The main novelties of our approach are in the computation of displacement uncertainty information and in using that information to support the outlier analysis. With this computationally inexpensive solution, a small set of features can provide estimates having reasonable accuracy in various scenes and situations.

2. LOCAL MOTION ANALYSIS

We consider first the computation of block motion estimates and the related uncertainty analysis (second step in Fig. 1).

2.1 Matching measure and displacement estimate

As a matching measure, our choice is to use the zero mean sum of squared differences (ZSSD) criterion, which can be defined as

$$D(d) = \sum_{x \in B} (I_1(x + d) - I_0(x) - \mu(d))^2,$$

where \(d = [u,v]^T\) is the displacement, \(B\) denotes a set of block pixel coordinates, \(I_0(\cdot)\) and \(I_1(\cdot)\) are the anchor and target frames, respectively, and \(\mu(d)\) is the average of \(I_1(x + d) - I_0(x)\) computed over \(x \in B\). Experimental results discussed in [3] show that this measure can tolerate changes in lighting conditions and sensor sensitivity. In our method the ZSSD is evaluated for some range of displacements. The surface of a matching measure values obtained in this way is called the motion profile.

The measure (1) is evaluated for discrete displacements whose components \(u\) and \(v\) are integers. The original estimate for the displacement, a minimizer of \(D(d)\), can be refined to subpixel precision by fitting of quadratic functions to criterion values in the neighborhood of that minimum (see [4, p.317]). Separate fitting in horizontal \((u)\) and vertical \((v)\) directions is performed in our case, and the locations of minima provide the estimate. According to our observation, the average difference between the estimated and actual displacements is reduced from 0.35 to 0.20 pixels with refinement.

2.2 Evaluating uncertainty of matching results

Using some difference measure such as ZSSD is based on an assumption that the actual displacement is close to the displacement that minimizes the criterion. However, there can be many displacements that give the minimum value or a value close to it, and the correct estimate might be close to any of them. For example, if a block is located on a straight edge, similarly textured blocks can be observed along the edge, which leads to multiple good matches.

As a basis for evaluating the confidence, we consider how the selection of good matches could be made. It has been noted that...
displaced frame differences are driven by local spatial intensity gradients in addition to the residual motion [5]. Based on this observation, it was proposed in [6] that simple gradient measures evaluated over block pixel values can be used for predicting how large values a block difference measure can have for a displacement candidate close to the true displacement.

This leads to a thresholding scheme where the set of good matches in the motion profile is defined as \( V = \{ d | D(d) \leq T \} \). The threshold \( T \) is computed using

\[
T = D(d_{\text{ref}}) + k_1 G + k_2
\]

where \( d_{\text{ref}} \) is a displacement which minimizes the block difference measure, \( G \) is a block gradient measure, and \( k_1 \) and \( k_2 \) are constants. The gradient measure is based on a sum of neighboring pixel differences and is defined as

\[
G = \sum_{x \in B_t} g^2(x, u_h) + \sum_{x \in B_v} g^2(x, u_v)
\]

where \( g(x, u) = (I_0(x + u) - I_0(x)) \), \( u_h = [1, 0]^T \), \( u_v = [0, 1]^T \), \( B_t \) is the block \( B \) extended with the row of pixels above it, and \( B_v \) is \( B \) extended with the columns of pixels left to it. Empirical justifications for using this rule and the choice of constants are discussed in Sec. 2.3.

Once thresholding has been performed, we summarize the result as a covariance matrix \( C_k \), which is the second central moment of \( d \in V \) with constant \( c = 1/12 \) added to the diagonal values. The constant reflects the fact that thresholding was done for integer-valued displacements. We also compute the first moment over \( V \), which is denoted as \( d_{\text{ref}} \) in the following.

Uncertainty about the displacement encoded in these moments can be represented in terms of the Mahalanobis distance between the vectors \( x_1 \) and \( x_2 \).

\[
d_{Mab}(x_1, x_2; \Sigma) = \sqrt{(x_1 - x_2)^T \Sigma^{-1} (x_1 - x_2)},
\]

where \( \Sigma \) is a covariance matrix. Any displacement \( d \) for which \( d_{Mab}(d, d; C_k) \) is below some proper threshold, is considered as a possible displacement. The overall analysis is illustrated in Fig. 2.

For our analysis, we divide the range of \( G \) to intervals and group samples according to those intervals. In each interval, we consider the probability \( P(k_1) \) of the event

\[
E(k_1) : \frac{D_c - D_m}{G + k_2} < k_1,
\]

where \( k_1 \) is a constant. Values of \( k_1 \), which correspond to specific probabilities \( P(k_1) \), are plotted as a function of \( G \) in Fig. 3 for two values of \( k_2 \). We note that for \( k_1 \approx 0.2 \), the probability \( P(k_1) \) is about 0.99. For low values of \( G \), the probability can be tuned by changing the value of \( k_2 \). An important thing is that \( P(k_1) \) does not seem to depend much on \( G \), and because the event \( E(k_1) \) can be restated as a thresholding rule used in the computation of \( V \) (let \( k_2 = k_1 k_2' \)), we have grounds for using the rule.

![Figure 3: The coefficient \( k_1 \) as a function of \( G \) for three different probabilities \( P(k_1) \) and two choices of \( k_2 \).](image)

If the values of \( k_1 \) and \( k_2 \) are increased, it becomes more and more probable that matches not close to \( d_{\text{ref}} \) become members of \( V \). Thus, \( k_1 \) and \( k_2 \) should not be arbitrarily high. To study this issue, we consider what kind of classification is in general made for those three displacements which are farthest away from \( d_{\text{ref}} \). (Fig. 4). For classification of displacements, we expect that

1. For any value of a gradient measure \( G \), the probability of including the best candidate to \( V \), denoted \( P_{\text{best}}(G) \) in the following, should be close to one.
2. A low value of \( G \) indicates that an image block is relatively homogeneous. Therefore, it is reasonable that the probability of excluding remote candidates from \( V \), denoted as \( P_{\text{ex}}(G) \), is low in those situations.
3. For higher values of \( G \), \( P_{\text{ex}}(G) \) should be high.

![Figure 4: Displacement candidates evaluated in the simulation based examination of thresholding performance.](image)

Probabilities \( P_{\text{best}}(G) \) and \( P_{\text{ex}}(G) \) are shown in Fig. 5 for some choices of \( k_3 \) when \( k_1 = 0.2 \). The probability \( P_{\text{ex}}(G) \) behaves as expected from the analysis above. With \( k_3 = 0 \), \( P_{\text{ex}}(G) \) is above zero for low values of \( G \), and can be pushed to zero by increasing \( k_2 \). For higher values of \( G \), remote candidates are not included to \( V \) in 90% of cases.

![Figure 5: Probabilities \( P_{\text{best}}(G) \) and \( P_{\text{ex}}(G) \) for different values of \( k_3 \).](image)
analysis. Several such features are computed for the purpose of the global motion estimation.

Locations p are obtained by dividing the image region to N rectangular subregions and taking one block from each subregion. In this way, we have features distributed over the image. Comparison of blocks in a subregion is based on analysis of the spatial gradient. For example, measures based on the Harris response function [7] or eigenvalues of the normal matrix [8] can be used. The latter approach is used in the experiments for this paper. In a subregion, we try to locate a block with strong gradient values in orthogonal directions first (both eigenvalues of the normal matrix are large). If such a block is not found, we seek a block with strong gradients in one spatial direction (large value for one eigenvalue).

3. GLOBAL MOTION ESTIMATION

Global motion estimation takes the local motion features and determines a parametric motion model, which is supported by those features. In general, such a model represents the displacement dh at the image coordinate p = [x, y]T as

\[ dh = H[p] \theta, \]  

where \( \theta \) is a N × 1 parameter vector and \( H[p] \) is a 2 × N mapping matrix which depends on p. For example, the four-parameter similarity model used in our experimental work has the mapping matrix

\[ H[p] = \begin{bmatrix} 1 & 0 & x & y \\ 0 & 1 & y & -x \end{bmatrix}. \]

The robust fitting of parametric models is typically based on the random sample consensus (RANSAC) method [9], or some variant of it. In our case, we present an approach where agreement on some model takes local motion uncertainty information into account. Recalling Fig. 1, the inlier selection takes the set of motion features as an input and outputs a set of inlier features which are triplets \((p, \hat{d}, W)\), where \( W \) is a 2 × 2 weighting matrix. This information is fed to the parametric model fitting stage, where the final result is estimated using the weighted least squares method [10].

3.1 Outlier analysis

In the outlier analysis, hypotheses about the global motion \( \theta \) are generated and voted for. To make up a hypothesis, a minimal subset of motion features is used to establish a system of equations based on (4), where \( \hat{d} \) from motion features are substituted for \( d_{inh} \). The best hypothesis is found among those generated, and it then defines the features to be used in the model fitting stage.

Selection of the best hypothesis is based on two counts of votes. The primary criterion is based on the notion that the hypothesis should induce local displacements which are possible according to the results of uncertainty analysis. The number of supporting features must be maximal. The secondary criterion considers local motion estimates \( \hat{d} \) and requires that the best hypothesis should induce the maximal number of displacements which are close to those estimates.

3.1.1 Primary criterion: uncertainty support

According to the results of the uncertainty analysis encoded in \( (d_{inh}, C_v) \), the hypothesis about the local displacement, \( \hat{d} \), is supported if

\[ d_{Mah}(d_{inh}, \hat{d}; C_v) \leq T_p, \]  

where \( T_p \) is some threshold for the distance. We compute the number of motion features where this condition holds, and this is our primary criterion for selecting a particular hypothesis.

The threshold \( T_p \) should be at least 2, which can be seen from the examples of the thresholded motion profiles shown in Fig. 6. We note that ellipses induced by \((d_{inh}, C_v)\) approximately cover candidates in \( V \) with this choice.

Figure 6: Patterns of thresholded motion profiles. Each black square corresponds to a motion candidate included in \( V \). Eigenvalues of \( C_v \) are given below the images. The shown ellipses correspond to Mahalanobis distances \( d_{Mah}(\cdot, \cdot; C_v) \in (0.5, 1, 1.5, 2) \).

3.1.2 Secondary criterion: estimate support

The secondary criterion for hypothesis selection considers the difference between \( d_{inh} \) and the estimate \( \hat{d} \). The idea is that if the difference is small, and the hypothesis is close to the true motion, then the local motion estimate provides information about the global motion and can therefore be used for the final parametric model fitting. The more such features there are, the better the hypothesis is.

When evaluating distances between \( d_{inh} \) and \( \hat{d} \), we classify first the motion features as either strong or weak ones. A strong feature is one which provides information about the local displacement at least in one spatial direction according to the results of the uncertainty analysis. A reasonable way to evaluate this is to consider the smaller eigenvalue, \( \lambda_2 \), of \( C_v \); we say that a motion feature is strong if

\[ \lambda_2 < T_s, \]

where \( T_s \) is a fixed threshold.

To choose a proper threshold \( T_s \), let us consider again the profile thresholding patterns shown in Fig. 6. Based on inspection, we can say that \( T_s \) must be at least 0.34, as then the pattern consisting of four displacements in a square (bottom-left image) is classified as a strong one. The reason for this pattern being an important one is that it plausibly occurs in some situations where the actual displacement is between those four displacements. In practice, we use a slightly larger value, \( T_s = 0.4 \).

Now, if a feature is a strong one, it accumulates the secondary criterion, if

\[ d_{Mah}(d_{inh} - \hat{d}; C_v) \leq T_p. \]

To understand the meaning of this rule, consider the representation of \((d_{inh} - \hat{d})\) as a sum \((e_1 + e_2)\), where \(e_1\) and \(e_2\) are parallel to the
eigenvectors corresponding to the maximum and minimum eigenvalues of \( C_2 \), respectively. The rule allows motion features, which carry only partial information about local motion to give a vote, because \( e_1 \) is allowed to be large.

If a feature is weak (uncertain), we cannot allow large differences, and the Mahalanobis distance cannot be used. However, if the difference is small in a Euclidean sense, it is probable that the displacement estimate can provide information about global motion in the fitting stage. Therefore, a weak feature gives a vote in the secondary criterion, if

\[
d_{\text{Eucl}}(d_h, \hat{d}) \leq T_e.
\]

where \( d_{\text{Eucl}} \) denotes the Euclidean distance between vectors, and \( T_e \) is a fixed distance value. In experiments, we use \( T_e = 1.0 \).

### 3.2 Weighting in model fitting

The parametric global motion model is fitted to displacement information of strong and weak features that vote in the secondary criterion for the best hypothesis. As noted above, the result is obtained using weighted least squares fitting. Our interest here is in how the \( 2 \times 2 \) weighting matrix should be selected for a particular feature. In a case of a strong feature, it is reasonable to use \( W = C_2^{-1} \) as \( C_1 \) encodes information about the displacement uncertainty.

With a weak feature, it is not reasonable to use \( C_1 \) as a weighting matrix. The idea is to use a weighting \( W \), for which

\[
d_{\text{Mah}}(d_h, \hat{d}; W^{-1}) \leq T_p
\]

is equivalent to rule (7). In this way, weighting of strong and weak features has a common basis. The weighting matrix becomes \( W = (T_p/T_e)^2I \). Note that increasing \( T_p \) or decreasing \( T_e \) gives more weight to a weak feature.

The outlier analysis and model fitting are illustrated in Fig. 7. Ellipses in the upper right image illustrate the results of the uncertainty analysis, whereas ellipses in the lower left image show the weighting for features.

![Figure 7: Example of global motion estimation](image)

#### 4. EXPERIMENTS

In the experiments, we evaluate the usefulness of the uncertainty analysis for global motion estimation. This is done by comparing the results to those obtained with an outlier analysis, where uncertainty information is not used, and inlier selection is just based on the Euclidean distance between the hypothesized and estimated displacements. Another interest in the experiments is to compare results obtained with our approach to those obtained with other methods. For this purpose, we estimate global motion using a multiresolution gradient based technique [1]. In the legends of figures, abbreviations FB and FB+UA refer to feature based motion estimation without and with uncertainty analysis, respectively, and GBM refers to the gradient based reference method. In our method, 16 features were used, the block size was \( 8 \times 8 \), the range of displacements was 16 pixels, and the thresholding parameters were \( k_1 = 0.2 \) and \( k_2 = 300 \).

Synthesized and real image sequences with a frame size of 160 by 120 pixels were used (Fig. 8). Synthesized sequences contain motion, which can be represented exactly with the similarity motion model. Real image sequences were taken with three different mobile phones, and color frames were converted to 8-bit luminance images. The sequences were annotated manually in order to obtain parametric approximations \( \theta \) of the ‘ground truth’ global motion.

![Figure 8: Sample frames from test sequences](image)

As a performance measure, the root mean square error (RMSE) of the estimates \( \hat{\theta} \) is evaluated. The RMSE is defined as the square root of the average value of \( \|H[p](\hat{\theta} - \theta)\|^2 \) computed over the image coordinates \( p \). For each method, RMSE values are obtained for all frame pairs in a set of image sequences, and the results are sorted in ascending order. It is then easy to see in corresponding plots, if there is any difference in performances. RMSE values over about 2-4 pixels indicate problems in the motion estimation.

#### 4.1 Synthetic sequences

In the case of synthesized sequences, we used images of 6 scenes to generate 18 sequences where the global motion can be accurately modelled using the similarity model. The length of each sequence was 31 frames. To test the robustness of methods, an independently moving outlier object was added to synthesized sequences (see Fig. 8a). In Fig. 9, the result of the experiments is illustrated for cases with and without it.

![Figure 9: Results with synthesized sequences](image)
based approach has an accuracy of about 0.1 pixels. As matching based methods cannot determine reliably subpixel displacements [11], such a performance can be expected. However, accurate fitting is typically not possible in real situations, and this reduces its importance.

As an indication of the usefulness of uncertainty analysis, we see from Fig. 9 that it improves the robustness of the feature based estimation. The performance achieved is not so high as with the gradient based method.

4.2 Scene image sequences

Scene image sequences were taken with two smartphones, Nokia 3650 (12 bits per color pixel) and N93 (24 bpp) phones. Especially in the former case, the sequences are noisy due to quantization (see Fig. 8b). The motion estimation results with those sequences are illustrated in Fig. 10.

One can see that the difference in performance of feature and gradient based methods that was observed with synthetic sequences does not exist with these sequences. With the ‘3650’ sequences, the result with the feature based method is slightly better, as large errors are avoided. It can be seen from Fig. 10 that uncertainty analysis improves the robustness of the approach.

4.3 Document scanning data

The method described in this paper has been applied in a document scanning application, where partial, slightly overlapping 640x480 images (e.g. 20%) of a document page are taken and used for composing an image of the whole page. Here, the purpose of the motion estimation method is to assist in the picture taking process. When displacement since the latest stored image goes above some threshold, and the user has stopped moving the camera, a new picture is taken. To detect these conditions, global motion is estimated between several consecutive low resolution (160x120) images. This provides cumulative displacement estimates, and the requirement is that the error in this estimate does not become too high in order to guarantee sufficient overlap between stored pictures, which are mosaicked later off-line.

For the experiments here, 10 sequences having a length of 20-50 frames were captured using a Nokia 6670 mobile phone (16 bpp). As can be seen from Fig. 11, the feature based method provides acceptable results with these sequences, whereas the gradient based method fails in about 20% of frame pairs. One reason for that might be that lines of text are similar in low resolution images. Using uncertainty information does not provide any benefit. The success of estimation must be due to the robustness of the ZSSD in dealing with illumination changes, which happen often in the sequences. Also using appropriate block selection criteria helps. The largest errors for cumulative displacement estimates were about five pixels, which is sufficient for the scanning application.

5. CONCLUSIONS

We have developed a block matching approach for global motion estimation which computes and utilizes uncertainty information about local displacements. The performance of the method has been characterized in experiments using both synthetic and real image sequences. With uncertainty analysis, large estimation errors can be reduced. How large the benefit is depends on the content of the imaged scene, but as a conclusion we can say that applicability of the approach is enhanced.

The method has been implemented as a software library and used in the development of smartphone applications that use video based motion estimation as a control input. Relatively low computational complexity has allowed us to do that.

One possible direction for future work is to reconsider outlier analysis from the viewpoint of the parametric model approximation quality. If a parametric model cannot model well the motion, then it may happen that only a small number of features can support some hypothesis. Some adaptive thresholding technique for outlier analysis might be used to relieve the problem.

REFERENCES