ROBUST OPTICAL FLOW ESTIMATION IN MPEG SEQUENCES

Konstantinos Rapantzikos, student member IEEE
School of Electrical & Computer Engineering
National Technical University of Athens
e-mail: rap@image.ntua.gr

Michalis Zervakis, member IEEE
Department of Electronic & Computer Engineering
Technical University of Crete
e-mail: michalis@systems.tuc.gr

ABSTRACT

Motion information is essential in many computer vision and video analysis tasks. Since MPEG is still one of the most prevalent formats for representing, transferring and storing video data, the analysis of its motion field is important for real time video indexing and segmentation, event analysis and surveillance applications. Our work considers the problem of improving the optical flow field in MPEG sequences. We address the issues of robust, incremental, dense optical flow estimation by combining information from two different velocity fields: the available MPEG motion field and the one inferred by a multiresolution robust regularization technique applied on the DC coefficients. Thus, the regularization technique is based only on information that is directly available in the compressed stream avoiding therefore the time and memory consuming decompression. We extend standard techniques by adding a temporal continuity and an MPEG consistency constraint, both as mathematical constraints in the objective function and as hypothesis tests for the presence of motion discontinuities. Our approach is shown to perform well over a range of different motion scenarios and can serve as a basis for efficient video analysis tasks.

1. INTRODUCTION

Analyzing motion patterns is essential for understanding visual surroundings. Representative applications of motion analysis include video interpolation, coding and transcoding, robotic vision, video indexing, medical imaging and super-resolution reconstruction [3, 6]. Direct processing in the compressed domain is essential due to the huge amount of encoded digital data available today [7, 8]. Nevertheless, several limitations are posed: Low-detail information due to lower resolution, oversmoothed spatial and motion structures, vaguely defined spatial and motion borders, intrinsic errors in the MPEG motion vectors.

The usual starting point for gradient-based velocity estimation is to assume that the intensities are shifted (locally translated) from one frame to the next and that the shifted intensity values are conserved (data conservation principle), i.e.

$$I(x, y, t) \approx I(x + u\delta t, y + v\delta t, t + \delta t) \quad (1)$$

where u, v denote the horizontal and vertical optical flow vector components and δt is small. This constraint implies that the intensity of a moving point in the image plane remains constant along the trajectory of the point in time. Gradient based methods use a linear approximation of Eq. (1) and obtain the *Optical Flow Constraint Equation* (OFCE)

$$I_x u + I_y v + I_t = 0 \quad (2)$$

where I_x, I_y indicate partial derivatives of the brightness function with respect to x and y and I_t indicates the partial derivative over time. The *generalized aperture problem* [1] that occurs mainly due to lack of sufficient intensity variation within local regions is handled by the addition of a spatial coherence assumption in the form of a regularizing term. Generally, the challenge is to achieve high robustness against strong assumption violations commonly met in real sequences. Several authors propose an extension of conventional techniques using robust statistics [1, 4]. We focus on the work of Black & Anandan, [1], and its extension towards dense optical flow recovery in compressed video.

We are dealing with the efficient combination of gradient-based and block matching motion estimation techniques under a single robust regularization framework to allow the generation of an improved motion field for an MPEG stream. In our approach we attempt to develop an efficient method that combines only their advantages over these regions in order to recover the true underlying motion as correct as possible. The novelty of our approach focuses on the fusion of available (MPEG) and generated (gradient-based) motion information and the use of new constraints on the motion field in the form of regularizing factors in the objective function.

2. DENSE ESTIMATION OF OPTICAL FLOW IN THE COMPRESSED DOMAIN

We start the presentation of our approach by realizing that most velocity estimation algorithms suffer from the initial value problem. Most optimization techniques fail, if the initial estimate of the solution is far from optimal. The MPEG standard provides valuable motion information that can be used to provide a "good" initial solution, namely the encoded MPEG motion vectors. Although not accurate, especially at the motion borders and homogenous regions, these motion vectors are something "more" than a crude initial motion field. To be used in this framework, the encoded MPEG motion vectors are transformed to a unified backward-predicted reference (section 2.1.1). In order to avoid full decompression of the MPEG stream we apply our robust estimation technique to the extracted DC images (section 2.1.1).

The exclusive operation in the compressed domain introduces several limitations. The low intensity resolution (DC block level) provides low detail information and produces spatially oversmoothed regions. Additionally, the intrinsic errors in the available MPEG motion vectors generate noisy areas and vaguely defined motion borders in the motion field. Thus, we need to incorporate additional motion constraints in order to compensate for these inefficiencies.

2.1 MPEG Information Extraction

The MPEG stream carries both motion and intensity information of the underlying scene. Motion is represented by a field of motion vectors and intensity is encoded into a set of Discrete Cosine Transform (DCT) coefficients. Processing in the compressed domain reduces the amount of effort involved in full decompression and keeps the storage cost low.

DCT coefficients are readily accessible for I frames, but they must be estimated for P and B frames. In essence, the DCT coefficients of the 16×16 macroblock (MB) area of the reference frame that the current P or B block was predicted from need to be calculated. Since the DCT is a linear transform, the DCT coefficients of the corresponding MB in the reference frame can be calculated from the four neighbouring MBs that overlap this reference MB, albeit with substantial computational expense. We incorporate the technique proposed by Yeo & Liu, [5] to calculate reasonable approximations to the DC coefficients of a MB of a P or B frame.

The MPEG frames may be of different types, i.e. I (no motion information), P (forward predicted) or B (backward prediced), and can occur in a variety of GOP (Group Of Pictures) patterns. An I frame has no motion vectors assigned to it in contrast to P (max one MV for every MB) or B (max two MVs for every MB) frames. We adopt the approach of Kobla *et al.* [2] to produce a unified-reference set of motion vectors that is independent of the frame type and the direction of prediction. This method represents each motion vector as a backward predicted vector with respect to the next frame, independently of frame type.

2.2 Construction of the Objective Function

Under the framework of our approach, we view the formulation and solution of the OFCE in relation to constraints provided by the available unified reference MPEG field. Motivated by a joint consideration of Bayesian and regularization approaches, we formulate our objective function as the combination of *observed* (data term) and *a priori* (smoothness constraints) information. Through the smoothness constraints we incorporate prior information regarding the spatial and temporal distribution of the estimated motion field, as well as the influence of the available MPEG motion field. Overall, the proposed objective function is formulated as:

$$E(\mathbf{u}) = \lambda_D E_D(\mathbf{u}) + \lambda_S E_S(\mathbf{u}) + \lambda_T E_T(\mathbf{u},\mathbf{u}^-) + \lambda_M E_M(\mathbf{u},\mathbf{u}_M) \quad (3)$$

initialized at the unified reference MPEG field, where $\mathbf{u}=[u,v]$, $\mathbf{u}^-$ is the previous velocity estimate λ_i with $i=\{D,S,T,M\}$ are weight factors and *D, S, T, M* stand for *data*, *smoothness*, *temporal* and *MPEG* indices, respectively. We use a Lorentzian robust error function $\rho(x,\sigma)$, as in [1], to resist against outliers yielding the following formulas for each separate energy part of Eq. (3):

$$E_D(\mathbf{u}) = \lambda_D \sum_{(x,y)\in\Re} \rho(I_x u + I_y v + I_t, \sigma_D)$$

$$E_S(\mathbf{u}) = \lambda_S \sum_{n\in G_s} \left[\rho(u-u_n,\sigma_S) + \rho(v-v_n,\sigma_S)\right]$$

$$E_T(\mathbf{u},\mathbf{u}^-) = \lambda_T \rho(u-u^-,\sigma_T) + \rho(v-v^-,\sigma_T)$$

$$E_M(\mathbf{u},\mathbf{u}_M) = \lambda_M \rho(u-u_M,\sigma_M) + \rho(v-v_M,\sigma_M)$$

where σ_i are the scale parameters of the robust estimator and G_s represents the north, south, east and west neighbors of n in the grid. Given this robust formulation, many optimisation techniques can be employed to recover the motion estimates. We use Simultaneous Over-Relaxation (SOR) to find the local minima and Graduated Non-Convexity (GNC) to find a globally optimal solution. The general idea is to take the non-convex objective function and construct a convex approximation. In the case of the Lorentzian estimator, this can be achieved by making the scale parameters ($\sigma_D, \sigma_S, \sigma_T$) sufficiently large. This approximation is then minimized using a coarse-to-fine (multiresolution) SOR technique. Successively better approximations of the true objective function are then constructed by altering the σ values, and minimized starting from the solution of the previous approximation. The multiresolution scheme of SOR makes the handling of large displacement in the scene more effective.

The individual constraints can be adaptively tuned, through their individual weights. We a selective (on-off) combination of the different terms in (3) that leads to computationally fast results, while retaining adequate accuracy.

2.3 Constraint-Weight Selection

The combination of the two motion fields, referring to the generated optical flow and the MPEG motion field, should take advantage of their competing nature in the overall criterion and improve the solution at the areas of their mismatch.

In order to employ the complementary nature of the motion fields under consideration and the advantages it offers, we design a method to balance the objective function's terms. It operates on the notion of inliers/outliers on the individual criteria, for updating and improving the motion field. We use two types of outliers, one referring to data coherence and the other to spatial smoothness, since they provide different information regarding the motion between two frames. Outliers are detected wherever the values of the data and spatial smoothness terms are greater than the outlier thresholds τ_D and τ_S, which in the case of a Lorentzian estimator are calculated as $\sqrt{2}\sigma_D$ and $\sqrt{2}\sigma_S$ respectively [1]. In this form, the consistency of a motion field with the differential OFCE can be readily rejected at the points of outliers in the data term (*data outliers*). In addition, discontinuities in the motion field can be detected at the points of outliers with respect to spatial smoothness (*spatial outliers*).

The motivation for using such a selective combination of constraints is to utilize the available information so as to 1) reduce the redundancy and the computational complexity of the algorithm and 2) to increase the effect of the MPEG motion field wherever it is "most likely" close to the true motion field. In this scheme, we enforce the MPEG at regions of motion discontinuities, as well as on smooth regions where the MPEG field does not violate the OFC. The decision about motion boundary and smooth regions is only activated when supported by both prior fields, namely the MPEG and previous (temporal) fields. It has to be mentioned here that the temporal motion field is supposed to be zero at the beginning. Moreover, in order to further assert the validity of the MPEG field, we initialize the

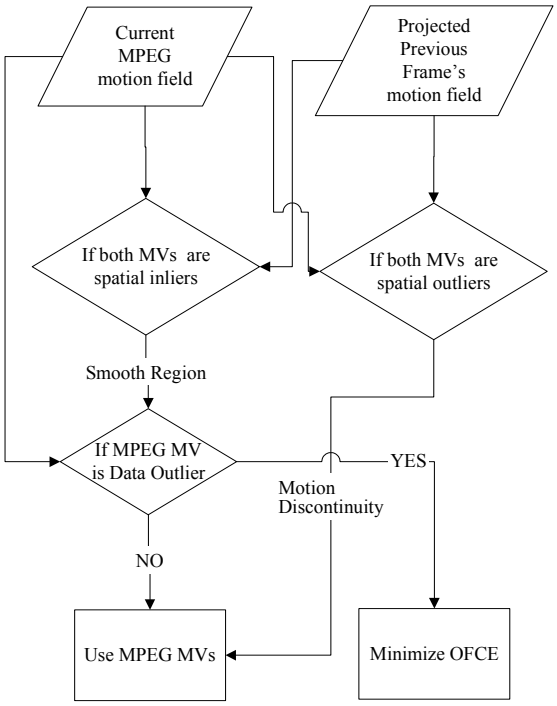

Fig. 1 Schematic diagram for weight decision

data constraint with the unified reference MPEG motion vectors that form the initial solution to our estimation approach. If the result is an inlier, then we may expect that the MPEG vector represents relative well the underlying motion, which means that it almost satisfies the OFCE. The weight selection scheme is illustrated in Fig. 1. A motion vector of the previous frame's motion field cannot be a spatial outlier. The case where the MPEG MV is spatial outlier and the previous MV is a spatial inlier is not depicted in Fig. 1 for simplicity. When the latter is the case we actually check if the MPEG MV is a data outlier or not and continue as shown in Fig. 1.

2.4 Scales Estimation

We use an automatic method for selecting the initial and final scales for the robust data conservation and smoothness constraints. As indicated before, the initial solution to our algorithm is the available MPEG motion field, which in a Bayesian framework can be viewed as *a priori* information. Assuming that most of these vectors are correct, i.e. fit well the data conservation term, we are based on them to obtain scale estimates. We initialize the OFCE with the MPEG motion vectors' components (u, v) and obtain a value for each pixel in the frame. We repeat this procedure for the smoothness constraint equation. The resulting distributions are assumed ε-contaminated Gaussian with means μ_i and standard deviations σ_i. This Gaussian assumption holds well for small residuals, which are located around the mean, but fails for large residuals forming the long tails of the distribution, which are due to wrong estimates of the MPEG vectors or to large inconsistencies between the matching and the OFC criteria. Therefore, we calculate an initial global scale σ_1 by fitting a Gaussian distribution, having in mind that this scale estimate can be "crude" and only used to generate the convex approximation of the objective function. Afterwards, we make a second fit on the residuals inside the interval $[-\sigma_1, +\sigma_1]$ and obtain a more accurate Gaussian fit with standard deviation σ_2. We then relate the σ_1 and σ_2 with the *max* and *min* scales of the Lorentzian functions used to model the data and smoothness terms. The attempted correspondence between Gaussian and Lorentzian scales can be justified by realizing that at small deviations the Lorentzian distribution approximates well the Gaussian

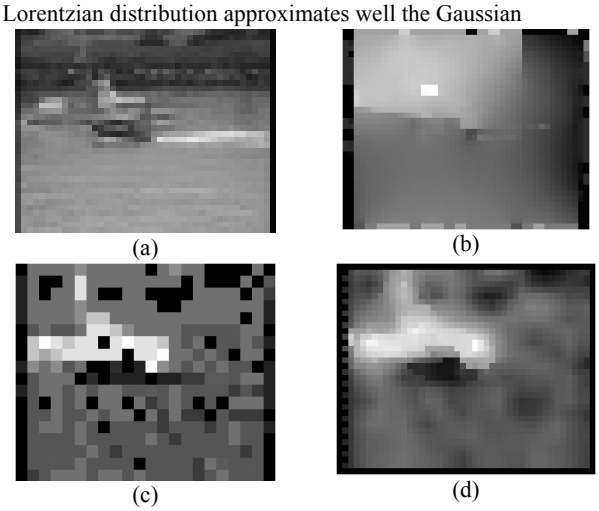

Fig. 2 (a) DC image, optical flow for (b) BA technique, (c) median filtered unified reference MPEG, (d) selective combination

The MPEG consistency and temporal continuity constraints impose a requirement to the derived motion field for being smoothly varying around the MPEG and temporal fields, respectively. Thus, the utilizing and consequently the structure of these constraints should resemble those of the smoothness constraint. Based on this reasoning, we use the same scales for the smoothness, MPEG and temporal constraints.

3. EXPERIMENTS

We processed several real video sequences to test the validity of our approach. We show only the obtained results for two well-known sequences. One MV per MB is calculated according to the objective function (2.2) and the rest are interpolated over the pixel grid using the nearest neighbor technique.

The *"coast guard"* shows a complex scene with different objects present. Fig. 2 depicts a zoomed region of a "coast guard" frame, showing the global optical flow field in terms of its velocity magnitude. The challenge here is to distinguish the boats without being affected by the global motion pattern (camera pan). The optical flow obtained by Black & Anandan's (BA) technique is overall smooth, as expected due to the camera pan. The non-informative temporal derivative computation due to the small temporal variation from frame to frame (short objects' motion) makes the boats indistinguishable. For comparison purposes, we further process the unified reference MPEG field by a vector median filter so as to remove spurious outliers. The filtered unified reference MPEG field provides a more distinct representation of the three motions present in the scene. It assigns almost zero velocities to the small boat so that

one can distinguish it from the global pan. The motion of the bigger boat is obvious. As expected, the MPEG field suffers from motion artifacts in homogenous regions, as illustrated by "gaps" and intensity discontinuities. The proposed approach provides similar results at the moving borders and seems to be more accurate. The global motion pattern is correctly assigned to the background, as shown by the smooth background areas.

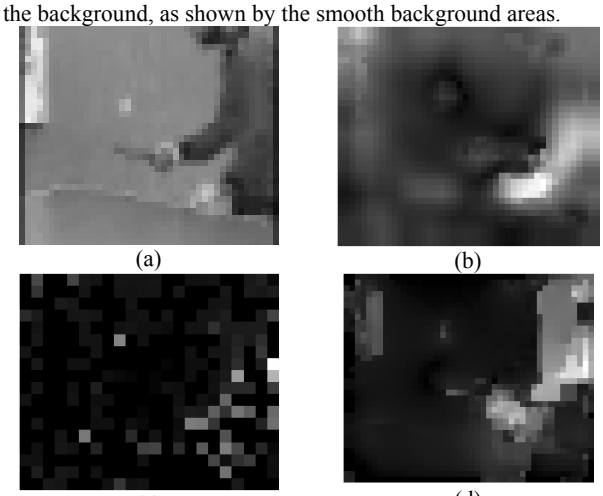

Fig. 3 (a) DC image; Optical flow for (b) our technique, (c) median filtered unified reference MPEG, (d) BA technique

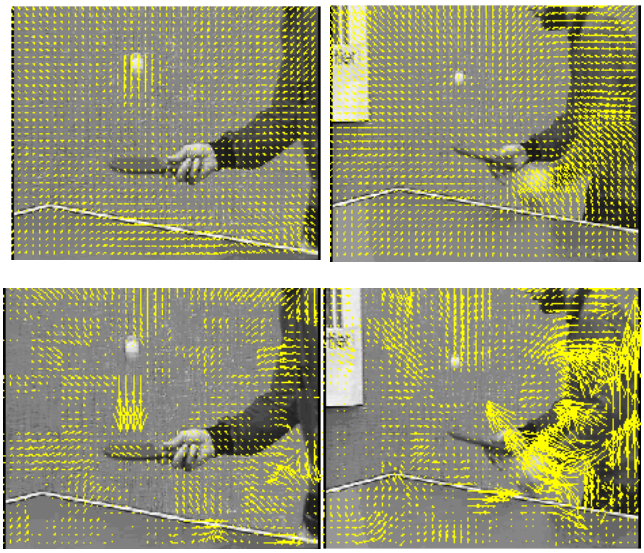

Fig 4 Least Squares blockwise fitting at frames 30 (first column) and 45 (second column)

Same conclusions are drawn from the "*table tennis*" sequence, as can be confirmed by the results in Fig. 3. In order to demonstrate and test the potential of the proposed method in motion characterization we further implemented a least squares regression algorithm on the last sequence. The flow is computed in a blockwise manner using a 6 parameter affine model:

$$\mathbf{u}(x, y; \mathbf{a}) = \begin{bmatrix} u(x,y) \\ v(x,y) \end{bmatrix} = \begin{bmatrix} a_0 + a_1 x + a_2 y \\ a_3 + a_4 x + a_5 y \end{bmatrix},$$

where **a** are the parameters of the model.

Fig. 4 illustrates the obtained results for two frames. The proposed method (first row) generates a smooth motion field and hence recovers correctly the global motion pattern. Although not with high accuracy, the different motions of the ball, bat and arm can be clearly distinguished. The same regression approach on the median filtered unified reference motion field (second row) presents many "gaps" due to false motion vectors and proves to be very noisy and inconsistent.

4. CONCLUSIONS

In this paper we introduce a framework for efficiently combining two motion estimation methods, compensating for their possible artifacts, and generating an improved, dense MPEG optical flow field. Our approach is limited from the quality of initial information we use, namely the DC images and the motion field. The use of DC images saves computational time due to the small spatial extend and the avoidance of decompression, but reflects strong smoothness in the intensity image. The combination of MPEG and temporal motion fields seems promising and is worth further analysis. Hard to face cases, like occlusion/disocclusion, illumination shading, appearance of new object etc. may be handled more efficiently with appropriate MPEG-temporal information fusion.

Acknowledgements

This work was supported by the EU STREP project: Improving airport Efficiency, Security and Passenger Flow by Enhanced Passenger Monitoring (OpTag-FP6-2002-Aero No. 502858).

5. REFERENCES

[1] Black, M., Anandan. P., "The Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow Fields", Computer Vision and Image understanding, vol. 63, no. 1, pp. 75-104, Jan 1996.

[2] Kobla, V., Doerman, D., Faloutsos. C., "Compressed Domain Video Indexing Techniques using DCT and Motion Information in MPEG Video", In Proc. Of the SPIE, vol. 3022, pp. 200-210, 1997.

[3] Leuven, J., Leeuwen, M.B. F.C.A. Groen., "*Real-Time Vehicle Tracking in Image Sequences",* In: *Proc. IEEE Instrumentation and Measurement Conference*, Budapest, Hungary, May 21-23, 2001, pp.2049-2054

[4] Memin, E., Perez, P, "*Dense estimation and object-based segmentation of the optical flow with robust techniques*", IEEE Trans. on Image Processing, Vol. 7, No. 5, pp. 703-719, May 1998

[5] Yeo, B.L., Liu., B., "On the Extraction of DC sequence from MPEG compressed Video", ICIP'95, pp. 260-263, 1995.

[6] Segall C.A., Molina R., Katsaggelos A.K., "High-Resolution Image from Low-Resolution Compressed Video", IEEE Signal Processing Magazine, pp. 37-48, May 2003.

[7] Benzougar A., Bouthemy P., Fablet R., "MRF-based moving object detection from MPEG coded video", ICIP'01, pp. 402-405, 2001.

[8] Tan Y,-P., Saur D.D., Kulkarni S. R., Ramadge P.J., "Rapid estimation of camera motion from compressed video with application to video annotation", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 1, Feb 2000.