Using Non-spatial Prior Information in Block-Matching Based Motion Estimation

Tarik Arici
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: tariq@ece.gatech.edu

Elif Albuz
Video Software Architecture Group
NVIDIA Corporation
2701 San Tomas Expressway
Santa Clara, CA 95050
Email: ealbuz@nvidia.com

Yucel Altunbasak
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: yucel@ece.gatech.edu

***Abstract*—Due to memory bandwidth limitations and computational complexity considerations in hardware implementations, block matching combined with ℓ_1 error norm and translational motion model is preferred in motion-estimation algorithms. Performance of this scheme is degraded by noise, compression artifacts, rotation, repeating structures, motion boundaries, zooming, and brightness changes. In this work, we present a Bayesian approach to incorporate prior information into block matching. Hypothesis testing is utilized to choose the most applicable prior motion vector and to compute a prior motion-vector distribution and its *precision*. Prior distribution is then updated with motion-vector likelihood derived from pixel data to obtain the posterior distribution, which is maximized via a search on the feasible motion-vector space.**

Keywords—**Motion estimation, block matching, prior motion information.**

I. Introduction

Motion estimation facilitates applications such as motion-compensated noise reduction, frame rate conversion, de-interlacing, and compression [1][2]. Block matching combined with translational-motion model and constancy of brightness assumption is preferred in hardware implementations. In block matching, the motion model applies to all pixels in a block, which simplifies memory access and resource requirements. To improve the performance, efficient buffering algorithms can be designed to fetch a block of pixel data in a small number of clock cycles [3]. However, when there is motion boundaries and/or deformation of objects, block matching can produce large errors in the motion-vector field. Translational-motion model reduces the computation and is sometimes even more robust to noise when compared to more complex models such as the affine motion model [4]. But it fails in the presence of rotation and zooming. After choosing the translational-motion model and its block based region of support, one needs to specify the estimation criteria of the model parameters (*i.e.*, x and y components of the motion vector). The constancy of brightness assumption tries to minimize the error between a pixel's intensity and its motion-compensated prediction's intensity. In hardware implementations, ℓ_1 norm is used in measuring the error magnitude. Compared to ℓ_2 norm, ℓ_1 norm is more robust in the presence of outliers and saves a multiplication operation [4][5]. ℓ_1 norm accumulated over all pixels in a block is called Sum of Absolute Deviations (SAD). SAD minimization is not sufficient to find true motion vectors, and performs poorly when the brightness in the scene changes.

Clearly, motion estimation is an ill-posed problem, which requires extra information other than pixel intensity data. Spatial and temporal correlation of the motion-vector fields can be used to regularize the motion estimation problem. Spatial correlation is induced because objects are usually larger than blocks [6]. Moreover, objects usually follow motion trajectories that does not abruptly change, which leads to temporal correlation. Tthe Bayesian framework is promising way to incorporate the prior information. Using Bayes law, a posterior probability for $p(v|d)$ of a realized motion-vector field v is computed by $p(d|v)p(v)$ up to a normalization constant, where $p(d|v)$ is the data likelihood, and $p(v)$ is the prior information.

Markov Random Fields (MRF) is a well-known method to impose spatial correlation. Maximum a posteriori (MAP) estimation of an MRF was introduced into computer vision by Geman and Geman [7]. The MAP-MRF framework can be expressed as an energy-minimization problem. Theoretically, it is possible to find the global minimum using simulated annealing, which is too slow to converge for practical purposes. Recently, approximation algorithms has been designed using graph cuts that iteratively updates the motion field [8][9]. Generally, energy-minimization based motion-estimation algorithms encode only the spatial correlation information via a spatial discontinuity-penalty term in the energy. The computational complexity of an energy-minimization problem that has temporal discontinuity will be too high, especially for hardware implementations: the temporal discontinuity-penalty term using the previous frame would require updating the previous frame's motion-vector field.

Spatial correlation by itself is not sufficient for creating a high-quality motion-vector field. One still needs to utilize the temporal correlation between the previous frame's already computed motion-vector field and the current frame's motion-vector field. By assuming independency, we rewrite $p(v)$ as $p^s(v)p^{\overline{s}}(v)$, where the two terms denote spatial and non-spatial prior motion information. $p^{\overline{s}}(v)$ encodes the temporal correlation between frames, and MAP estimate of $p(v|d)$ can be performed on the current motion-vector field. In addition we show how to use $p^{\overline{s}}(v)$ to pass information from previous resolutions in hierarchical motion estimation. Since, we do not update the previously computed motion-vector field of previous frame or previous resolution, we need to choose which motion vector to use in $p^{\overline{s}}(v)$ formulation. We present the prior motion-vector selection as a multi-hypothesis testing problem, and use the *evidence* for the winning hypothesis to adjust the precision[1]

[1]We refer to reliability of the prior as precision, following the convention for Gaussian distribution, which we use to model the prior distribution.

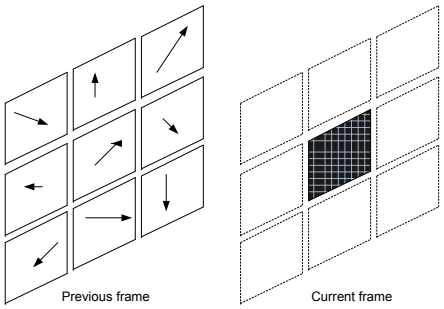


Fig. 1. Nine motion vectors from the previous frame are hypothesized for the block (shown as black) in the current frame)

of $p^{\overline{s}}(v)$[2].

II. Motion Estimation via Block Matching

In this section, we will present our Bayesian approach to incorporate non-spatial prior motion information into block matching.

Usually, there are more than one possible prior motion vectors (v_i^p) available (*e.g.*, see Figure 1)[3]. Unfortunately, we do not know which block's motion vector in the previous frame applies to the current block, because this is actually the motion estimation problem we are trying to solve. But we can expect that the current block must be a displaced version of one of the blocks that are not too far from its location in the previous frame. To choose the best v_i^p from a set of motion vectors, $\{v_i^p\}$, we use multiple hypothesis testing, which is described next.

A. Multiple Hypothesis Testing for Prior Selection

We have a set of hypotheses $\{H_1, H_2, \ldots, H_N\}$ to be tested. H_i hypothesizes that the motion vector v is equal to v_i^p from the previous frame. We want to find the hypothesis that has the highest *evidence*. The evidence value for H_i is defined as

$$e(H_i) = \ln \frac{p(H_i|d, v^p)}{p(\overline{H}_i|d, v^p)}, \tag{1}$$

where d denotes data (pixel intensities), v^p denotes prior information on previous frame's motion vectors, and $\overline{H}_i$ implies H_i is false [10]. $p(H_i|d, v^p)$ and $p(\overline{H}_i|d, v^p)$ are posterior probabilities of H_i and $\overline{H}_i$ obtained via Bayes Law:

$$p(H_i|d, v^p) = p(H_i|v^p)\frac{p(d|H_i, v^p)}{p(d|v^p)}, \tag{2}$$

$$p(\overline{H}_i|d, v^p) = p(\overline{H}_i|v^p)\frac{p(d|\overline{H}_i, v^p)}{p(d|v^p)}, \tag{3}$$

where $p(H_i|v^p)$ and $p(\overline{H}_i|v^p)$ denote prior information on H_i and $\overline{H}_i$, respectively.

Using (2) and (3) in (1)

$$e(H_i) = \ln \frac{p(H_i|v^p)p(d|H_i, v^p)}{p(\overline{H}_i|v^p)p(d|\overline{H}_i, v^p)} \tag{4}$$

By applications of Bayes Law on $p(d|\overline{H}_i, v^p)$, $e(H_i)$ becomes

$$e(H_i) = \ln \frac{p(d|H_i, v^p)p(H_i|v^p)}{\sum_{k=1, s.t. k\neq i}^{N} p(d|H_k, v^p)p(H_k|v^p)}. \tag{5}$$

[2]From now on, we will drop the superscript $\overline{s}$ for notational simplicity.

[3]Superscript p denotes data from the previous frame.

In the above formula, $p(H_i|v^p)$ represents our prior information on H_i before observing any data (*i.e.* the current frame). Generally, motion estimation algorithms output a motion vector and an associated confidence value c_i^p for that motion vector. Hence, before processing a new frame if we know that v_i^p has a high confidence, then it is more likely to be an estimate of a true motion in the image. This makes it more likely to survive in the new frame. Therefore, we can compute $p(H_i|v^p)$ by

$$p(H_i|v_p) = \frac{c_i^p}{\sum_{k=1}^{N} c_k^p}. \tag{6}$$

To select the best prior, we solve

$$i^* = \arg \max_{i\in\{1,2,\ldots,N\}} e(H_i). \tag{7}$$

However, if we do not use any prior information on H_i's by assuming $p(H_i|v^p)$ is uniform, (5) is simplified to

$$e(H_i) = \ln \frac{p(d|H_i, v^p)}{\sum_{k=1,}^{N} p(d|H_k, v^p) - p(d|H_i, v^p)}. \tag{8}$$

Since $e(H_i)$ in (8) is a strictly increasing function of $p(d|H_i, v^p)$, maximization problem in (7) becomes

$$i^* = \arg \max_{i\in\{1,2,\ldots,N\}} p(d|H_i, v^p), \tag{9}$$

which is simpler and does not involve division as in (7).

Although solving the simpler maximization in (9) gives the optimal i^* for the maximization in (7) by assuming uniform prior on H_i's, we still need to compute $e(H_i)$. This is because we use the top two largest evidence values, $e(H_{i^*})$ and $e(H_{i^{**}})$ to adjust the precision, $\frac{1}{\sigma^2}$, of $p(v)$ as below

$$\frac{1}{\sigma^2} = f(e(H_{i^*}) - e(H_{i^{**}})), \tag{10}$$

where f is a non-decreasing, non-negative function. A large difference between the two largest evidence values implies that the two hypothesis are well separated and we can be more certain that our decision of selecting H_{i^*} is right.

We would like to simplify $e(H_i)$ further by an approximation. When the evidence is large for H_{i^*} (*i.e.*, $p(d|H_{i^*}, v^p)$ is large), it should be saturated to avoid assigning too high $\frac{1}{\sigma^2}$. When the evidence is small, we want to adjust $\frac{1}{\sigma^2}$ by (10). Hence, our approximation of $e(H_i)$ should especially work well for small $p(d|H_i, v^p)$. For $\sum_{k=1,}^{N} p(d|H_k, v^p) \gg p(d|H_i, v^p)$, we can approximate $e(H_i)$ by

$$e(H_i) = \ln \frac{p(d|H_i, v^p)}{\sum_{k=1,}^{N} p(d|H_k, v^p)}. \tag{11}$$

Substituting (11) in (10), we get

$$\frac{1}{\sigma^2} = f(\ln p(d|H_{i^*}, v^p) - \ln p(d|H_{i^{**}}, v^p)), \tag{12}$$

which is the difference of the likelihoods of the top-two hypothesis.

Above formula for $\frac{1}{\sigma^2}$ measures if H_{i^*} is well separated from the rest of the hypothesis or not. However, the way we set our hypotheses can reduce the inference of a good H_{i^*}. Our H_i's are *simple* hypotheses meaning that each hypothesis specifies a single value to v (*i.e.*, $H_i : v = v_i^p$). This means that close but

not identical (*e.g.*, sub-pixel different) v_i^p's all of which applies well to the current frame, can be assigned to different H_i's. But from (11), $\frac{1}{\sigma^2}$ will be small since both H_{i^*} and $H_{i^{**}}$ will have high likelihoods. To overcome this, one can impose a minimum distance among the hypothesized v_i^p's.

One last improvement on our hypotheses set is to include a *dummy* hypothesis, H^D, to represent cases like occlusion and scene change in which no prior motion information is available for the next frame. Adding a *dummy* hypothesis enables us to choose H^D when evidence for v_i^p's are small. Obviously, the likelihood given H^D does not depend on pixel intensity differences, hence we set $p(d|H^D, v^p) = \epsilon$, where ϵ is a small number. When H^D is selected, we do not have any informative prior information $p(v)$ to update $p(d|v)$ because there is no match of pixels by the definition of H^D. Therefore, we need to set $\frac{1}{\sigma^2}$ to zero. To do this by smoothly changing $\frac{1}{\sigma^2}$ as H^D becomes more likely to be accepted, we can modify (10) to

$$\frac{1}{\sigma^2} = f[(\ln p(d|H_{i^*}, v^p) - \ln p(d|H_{i^{**}}, v^p))(\ln p(d|H_{i^*}, v^p) - \ln \epsilon)]. \quad (13)$$

As the likelihood $p(d|H_{i^*}, v^p)$ of the selected hypothesis H_{i^*} decreases (H^D becomes more plausible), the precision $\frac{1}{\sigma^2}$ of the prior information also decreases and becomes exactly zero when H^D is selected.

B. Computing the Posterior Distribution

After selecting the prior motion vector $v_{i^*}^p$ and computing its precision $\frac{1}{\sigma^2}$, we need to model the prior distribution $p(v)$. We expect the true motion vector to be *closely* distributed around $v_{i^*}^p$, and $\frac{1}{\sigma^2}$ hints how close this is. In a sense, $v_{i^*}^p$ and $\frac{1}{\sigma^2}$ specifies the first and the second moments of the prior distribution. To avoid imposing any further constraints on the prior distribution, Gaussian distribution is chosen to model $p(v)$ according to the principle of maximum entropy [10]

$$p(v) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(v - v_{i^*}^p)'(v - v_{i^*}^p)}, \quad (14)$$

which assumes horizontal and vertical deviations from v_i^p are not correlated.

The error between a pixel x's intensity and its reference pixel's intensity obtained by motion compensating with v is denoted by d_v^x, and it is created by many error sources such as noise, aliasing, compression artifacts, deformation, zooming, rotation, brightness change, etc. Again, by the principle of maximum entropy to account for all these sources Gaussian distribution is used for modeling data likelihood

$$p(d_v^x|v) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{1}{2\tau^2} {d_v^x}^2}, \quad (15)$$

where τ is a constant variance term reflecting the strength of the above mentioned error sources in the video sequence. Since block matching assumes the same motion vector v applies to all pixels in block B, the data likelihood for B is

$$p(d|v) = \prod_{x \in B} p(d_v^x|v), \quad (16)$$

$$= \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{1}{2\tau^2} \sum_{x \in B} {d_v^x}^2}, \quad (17)$$

by assuming d_v^x's are independently distributed.

From (17) and (14), the posterior distribution of v is

$$p(v|d) \propto p(d|v)p(v) \propto \frac{1}{2\pi\sigma\tau} e^{-\frac{1}{2\tau^2} \sum_{x \in B} {d_v^x}^2} e^{-\frac{1}{2\sigma^2}(v - v_{i^*}^p)'(v - v_{i^*}^p)} \quad (18)$$

The log of the posterior is

$$\log p(v|d) \propto -\frac{1}{2\tau^2} \sum_{x \in B} {d_v^x}^2 - \frac{1}{2\sigma^2}(v - v_{i^*}^p)'(v - v_{*i}^p) \quad (19)$$

MAP estimate, v^{MAP}, is found by minimizing $-\log p(v|d)$ given by

$$v^{MAP} = \arg\min_v \frac{1}{2\tau^2} \sum_{x \in B} {d_v^x}^2 + \frac{1}{2\sigma^2}(v - v_{i^*}^p)'(v - v_{*i}^p). \quad (20)$$

Substituting (17) in (13), $\frac{1}{\sigma^2}$ can also be simplified to

$$\frac{1}{\sigma^2} = f[(\frac{1}{2\tau^2} \sum_{x \in B} {d_{v_{i^{**}}^p}^x}^2 - \frac{1}{2\tau^2} \sum_{x \in B} {d_{v_{i^*}^p}^x}^2)(\ln \frac{1}{\epsilon} - \frac{1}{2\tau^2} \sum_{x \in B} {d_{v_{i^*}^p}^x}^2)]. \quad (21)$$

Inspecting (20), we can see that the MAP estimate minimizes a bi-criterion cost function. The first term is the data term, which is a sum of square errors (SSE). The second term penalizes deviations from the hypothesized motion vector. The weight of the second term is adjusted by how well the hypothesized motion vector applies to the current block B, which is inferred using (21). If the data likelihood and prior distributions are modeled with Laplace distribution instead of a Gaussian distribution, all the SSE terms in (20) and (21) would be SAD terms, which is more hardware friendly. Searching for the minimum of the first term alone corresponds to regular block matching that minimizes SAD.

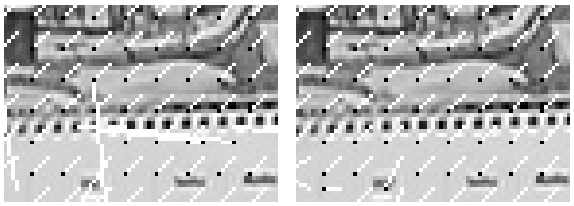
The search window, S, is exhaustively searched for finding v^{MAP}. Heuristics search algorithms such as three-step search [11], cross-search [12] for minimizing (20) can also be used instead of an exhaustive search. Furthermore, although the first term is a non-convex cost function, the second term is a convex cost function, which becomes more dominant as v's distant from $v_{i^*}^p$ are searched. This can be taken advantage of, while designing heuristic search algorithms specifically for minimizing (20). More costly searches such as full-search can be performed when non-convex cost function is dominant (*i.e.*, v's close to $v_{i^*}^p$) to avoid getting stuck at a local minimum, and heuristic search patterns can be utilized otherwise.

C. Choosing The Best Set of Hypotheses

To incorporate prior motion information from the previous frame, the hypothesis set should cover a large enough area in the previous frame that contains B's pixels while keeping the computation at an acceptable level. Obviously, this depends on the motion type. As shown in Figure 1, in addition to B's collocated block in the previous frame, its 3×3 block neighborhood can also be used. With the *dummy* hypothesis, this will make a total of ten hypotheses to utilize the available prior information from the previous frame. In hierarchical motion estimation, the hypothesis set needs to be designed differently. For example, if down-sampling by two is performed to create the coarser resolution frame from the finer resolution frame, B in the finer resolution will effectively correspond to a quarter-sized block in the coarser resolution. Hence, its motion information may be lost if there is a motion boundary in the whole block. To

(a) (b)

Fig. 2. Block matching results for a two-resolution hierarchical motion estimation: (a) SAD minimization (b) minimization of bi-criterion cost in Equation (20).



(a) (b)

Fig. 3. Block matching results for passing prior motion information from the previous frame: (a) SAD minimization (b) minimization of bi-criterion cost in Equation (20).

include this lost information in our hypothesis set, we need to use other neighbor blocks in the coarse resolution. For any B at location (r, c), corresponding to row and column number respectively, we need to use four blocks in the coarse resolution to account for all possible motion boundary directions. These four blocks are located at $(\lfloor \frac{r+1}{2} \rfloor, \lfloor \frac{c+1}{2} \rfloor)$, $(\lfloor \frac{r+1}{2} - 1 \rfloor, \lfloor \frac{c+1}{2} \rfloor)$, $(\lfloor \frac{r+1}{2} \rfloor, \lfloor \frac{c+1}{2} - 1 \rfloor)$, $(\lfloor \frac{r+1}{2} - 1 \rfloor, \lfloor \frac{c+1}{2} - 1 \rfloor)$. With the *dummy* hypothesis, there will be five hypotheses in our set.

III. Experiment Results

We tested our proposed method using standard video test sequences. In all the test sequences, using prior motion information has improved the quality of the motion-vector field. Due to space limitations, we present some visual results from the Mobile and Calendar sequence. In the following figures, motion vector of a block is represented with a white line and its direction is denoted by a block dot. Unfortunately, we can not do any mean square error (MSE) comparison by measuring the error between the original image and the reconstructed image via motion compensation. It is obvious from (20) that SAD minimization will always give a better MSE because of our second cost term. Therefore, we present visual results in which motion vectors are amplified by two.

Our first result demonstrates the improvement provided by passing prior motion information from a hierarchical motion estimation with two resolution levels. Figure 2(a), and 2(b) shows the motion-vector field from the 25^{th} frame, produced by SAD minimization and our proposed bi-criterion cost minimization in (20). The hypothesis set is as described in Section II-C. Due to noise in the image, there are some random motion vectors in Figure 2(a), which are worst in the smooth areas on the calendar. The motion-vector field in Figure 2(b) looks more consistent, most of the random vectors are corrected with the help of the prior information passed from the previous resolution.

The next result is produced by passing prior motion information from the previous frame. Figure 3(a) and Figure 3(b), shows the motion-vector field from the 22^{nd} frame produced by SAD minimization and our proposed method, respectively. Using the previous frame our proposed method improves the performance of block matching for repeating structures (*i.e.*, spiral calendar perforations).

It is important to note that, the quality of the motion-vector field still needs to be improved using an energy-minimization algorithm that imposes a spatial regularization as described in Section I.

References

[1] G. de Haan, *Video processing for multimedia systems*, Eindhoven, 2000.

[2] R. C. Gonzales and P. Wintz, *Digital Image Processing*, Addison-Wesley, Reading, MA, 1977.

[3] Aleksandar Beric, Ramanathan Sethuraman, Jef L. van Meerbergen, and Gerard de Haan, "Memory-centric motion estimator.," *IEEE Conference on VLSI Design*, pp. 816–819, 2005.

[4] J. Konrad, *Motion detection and estimation in Handbook of Image and Video Processing*, Academic Press, 2005.

[5] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004.

[6] G. de Haan and P.W.A.C. Biezen, "An efficient true-motion estimator using candidate vectors from a parametric motion model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 1, pp. 85–91, February 1998.

[7] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," vol. 6, no. 6, pp. 721–741, November 1984.

[8] V. Kolmogorov and Y.Y. Boykov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," in *EMMCVPR02*, 2002, p. 359 ff.

[9] Y.Y. Boykov and O. Veksler, "Graph cuts in vision and graphics: Theories and applications," 2005, pp. 79–96.

[10] E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, 2003.

[11] B. Zeng, R.X. Li, and M.L. Liou, "Optimization of fast block motion estimation algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 6, pp. 833–844, December 1997.

[12] R.X. Li, B. Zeng, and M.L. Liou, "A new three-step search algorithm for block motion estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 4, no. 4, pp. 438–442, August 1994.