Robust Global Motion Compensation in Presence of Predominant Foreground

Seyed Morteza Safdarnejad
https://www.msu.edu/~safdarne/
Xiaoming Liu
http://www.cse.msu.edu/~liuxm/
Lalita Udpa
http://www.egr.msu.edu/ndel/profile/lalita-udpa

Michigan State University
East Lansing
Michigan, USA

The objective of global motion compensation (GMC) is to remove *intentional* (due to camera pan/tilt/zoom) and *unwanted* (e.g., due to hand shaking) camera motion. GMC is utilized in applications such as video stitching, or as pre-processing for motion-based video analysis. Normally, GMC estimates the homography transformation between two consecutive frames by matching keypoints on the frames, and mapping the frames to a global coordinate. To remedy outliers in keypoint matches, robust techniques are proposed for homography estimation, e.g., RANSAC [1], by assuming the number of outliers to the correct homography is much less than inliers. However, in the presence of *predominant foreground*, i.e., moving objects and people, a larger proportion of the putative matches are mismatches. Predominant foreground may result from a higher percentage of coverage by foreground pixels, or occlusion, textureless and non-informative background, blurred background, or a combination of these reasons. In presence of predominant foreground, the common variations of RANSAC have little chance of selecting a minimal set of background keypoints by random sub-sampling in a limited number of iterations.

In this paper, we propose a robust GMC (RGMC) method for suppressing foreground keypoint matches and mismatches, enabling a reliable homography estimation in presence of predominant foreground and textureless background (Fig. 1). We perform *foreground suppression* by clustering motion vectors computed from keypoint matches and identifying potential clusters corresponding to the background. We use SURF algorithm for keypoint detection and description and to detect sufficient background keypoints even for textureless backgrounds, the Fast-Hessian keypoint detection threshold is decreased drastically. Since motion vectors on the background result from camera motion and are more *consistent* than foreground motion vectors, clustering will likely lead to some candidate regions from the background (see Fig. 1 (a)). Each cluster is analyzed separately by random subsampling of matches in that cluster and evaluating the resultant homography against the cost function, discussed later. However, background motion vectors may be assigned to multiple clusters. To merge background clusters, based on the estimated homography and cost function value (CFV) of each cluster, a subset of the best clusters are selected to be merged in a greedy algorithm (Fig. 1(b)).

To evaluate the estimated homography from a quadruplet of keypoints matches, we derive a cost function that unifies the keypoint matching score, edge matching score, and the information from compensating previous frames. Denote the matching frames as $\mathbf{I}_{t-1}$ and $\mathbf{I}_t$, their candidate homography as θ_t, and the set of keypoint matches under study as $\mathbf{D}$. In Bayesian framework, θ_t can be estimated by maximizing

$$p(\theta_t|\mathbf{D},\mathbf{I}_t,\mathbf{I}_{t-1},\theta_{t-1}) = \frac{p(\mathbf{D},\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_t,\theta_{t-1})p(\theta_t|\theta_{t-1})}{p(\mathbf{D},\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_{t-1})}, \quad (1)$$

where θ_{t-1} is the obtained prior homography of frames $\mathbf{I}_{t-1}$ and $\mathbf{I}_{t-2}$. The $p(\theta_t|\theta_{t-1})$ is the conditional probability of θ_t given the prior homography θ_{t-1}. The denominator of Eqn. 1 is constant w.r.t. θ_t. By expanding the likelihood term, the homography can be verified using

$$p(\theta_t|\mathbf{D},\mathbf{I}_t,\mathbf{I}_{t-1},\theta_{t-1}) \propto p(\mathbf{D}|\mathbf{I}_t,\mathbf{I}_{t-1},\theta_t,\theta_{t-1})p(\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_t,\theta_{t-1})p(\theta_t|\theta_{t-1}). \quad (2)$$

The term $p(\mathbf{D}|\mathbf{I}_t,\mathbf{I}_{t-1},\theta_t,\theta_{t-1}) = p(\mathbf{D}|\mathbf{I}_t,\mathbf{I}_{t-1},\theta_t)$ and represents how well the keypoint matches $\mathbf{D}$ extracted from $\mathbf{I}_t$ and $\mathbf{I}_{t-1}$ are matched by θ_t. Knowing $\mathbf{I}_t$ is independent from θ_{t-1}, $p(\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_t,\theta_{t-1}) = p(\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_t)$, reflects how well the frame $\mathbf{I}_t$ transformed under θ_t matches $\mathbf{I}_{t-1}$. Thus, the homography is estimated by minimizing,

$$f(\theta_t) = -\ln(p(\mathbf{D}|\mathbf{I}_t,\mathbf{I}_{t-1},\theta_t)) - \ln(p(\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_t)) - \ln(p(\theta_t|\theta_{t-1})). \quad (3)$$

We use the keypoint matching error formulation of Yan et al. [2] for $p(\mathbf{D}|\mathbf{I}_t,\mathbf{I}_{t-1},\theta_t)$. For appearance consistency under θ_t transformation,

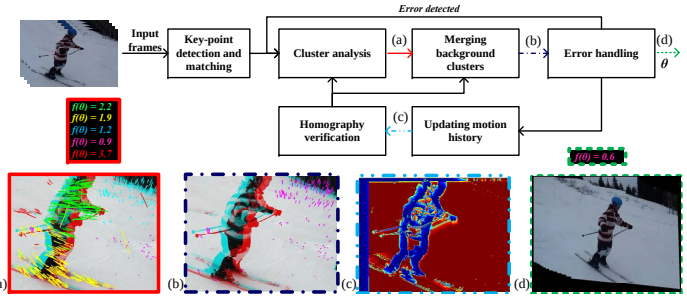

Figure 1: RGMC algorithm flowchart: (a) color indicates various motion vector clusters, (b) the merged cluster of background, (c) the motion history, and (d) the motion compensated video.

$p(\mathbf{I}_t,\mathbf{I}_{t-1}|\theta_t)$, we propose edge-based matching for its superior accuracy and efficiency.

To insure that the edge matching score reflects how well the background, not foreground, of the two frames match, we iteratively update a motion history $\mathbf{M}_t$. We define the edge matching score (EMS) as

$$E(\mathbf{I}_1,\mathbf{I}_2,\mathbf{R}) = \frac{2\|\Phi(\mathbf{I}_1)\bigodot\Phi(\mathbf{I}_2)\bigodot\mathbf{R}\|_1}{\|\Phi(\mathbf{I}_1)\bigodot\mathbf{R}\|_1 + \|\Phi(\mathbf{I}_2)\bigodot\mathbf{R}\|_1 + c}, \quad (4)$$

where Φ is edge detection operator, $\bigodot$ is element-wise multiplication, $\mathbf{R}$ denotes the mask specifying the region of interest for EMS calculation, $\|\cdot\|_1$ computes the L_1 matrix norm, and $c(=0.001)$ is a constant to avoid division by zero. We use thresholded motion history $\mathbf{M}$ as region of interest in our homography verification model.

To utilize the prior information for a stable homography estimation, we decompose the homography model into translation, scale, and rotation models and model the difference between these values over consecutive frames.

We empirically learn the various probability models used in our formulations. For this learning, we manually stitch 250 pairs of consecutive frames to find the best homography estimate. By plugging the probability models to Eqn. 3 and ignoring the constants, the final cost function is,

$$f(\theta_t) = \frac{\sum_{i=1}^{N_{in}}\frac{d_i}{b} + \sum_{i=1}^{N_{out}}\ln(S^2)}{N_{in}+N_{out}} + \frac{(E(\mathbf{I}_t,\mathbf{I}_{t-1},\mathbf{M})-\mu_E)^2}{2\sigma_E^2} + \lfloor\frac{(\Delta s-\mu_{\Delta s})^2}{2\sigma_{\Delta s}^2}\rfloor_T + \lfloor\frac{(\Delta\alpha-\mu_{\Delta\alpha})^2}{2\sigma_{\Delta\alpha}^2}\rfloor_T + \lfloor\frac{|\Delta t_x-\mu_{\Delta t_x}|}{b_{t_x}}\rfloor_T + \lfloor\frac{|\Delta t_y-\mu_{\Delta t_y}|}{b_{t_y}}\rfloor_T, \quad (5)$$

where N_{in} and N_{out} represent number of inliers and outliers, respectively, and $\lfloor x\rfloor_T = \min(x,T)$ restricts the impact of prior information. The homography θ_t is estimated by

$$\theta_t^* = \arg\min(f(\theta_t)). \quad (6)$$

Extensive experiments and comparison with manually matched ground truth and baseline methods demonstrate the superiority of RGMC.

[1] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *ACM Trans. Communications*, 24(6):381–395, 1981.

[2] Qing Yan, Yi Xu, Xiaokang Yang, and Truong Nguyen. Heask: Robust homography estimation based on appearance similarity and keypoint correspondences. *Pattern Recognition*, 47(1):368–387, 2014.