Temporally Robust Global Motion Compensation by Keypoint-Based Congealing

S. Morteza Safdarnejad(✉), Yousef Atoum, and Xiaoming Liu

Michigan State University, East Lansing, USA
safdarne@egr.msu.edu, atoumyou@msu.edu, liuxm@cse.msu.edu

Abstract. Global motion compensation (GMC) removes the impact of camera motion and creates a video in which the background appears static over the progression of time. Various vision problems, such as human activity recognition, background reconstruction, and multi-object tracking can benefit from GMC. Existing GMC algorithms rely on sequentially processing consecutive frames, by estimating the transformation mapping the two frames, and obtaining a composite transformation to a global motion compensated coordinate. Sequential GMC suffers from temporal drift of frames from the accurate global coordinate, due to either error accumulation or sporadic failures of motion estimation at a few frames. We propose a temporally robust global motion compensation (TRGMC) algorithm which performs accurate and stable GMC, despite complicated and long-term camera motion. TRGMC densely connects pairs of frames, by matching local keypoints of each frame. A joint alignment of these frames is formulated as a novel keypoint-based congealing problem, where the transformation of each frame is updated iteratively, such that the spatial coordinates for the start and end points of matched keypoints are identical. Experimental results demonstrate that TRGMC has superior performance in a wide range of scenarios.

Keywords: Global motion compensation · Congealing · Motion panorama

1 Introduction

Global motion compensation (GMC) removes the impact of *intentional* and *unwanted* camera motion in the video, transforming the video to have *static* background with the only motion coming from foreground objects. As a related problem, video stabilization removes *unwanted* camera motion, such as vibration, and generates a video with a *smooth* camera motion. The term "global motion compensation" is also used in video coding literature, where background motion is estimated roughly to enhance the video compression performance [1,2].

GMC is an essential module for processing videos from *non-stationary* cameras, which are abundant due to emerging mobile sensors, e.g., wearable cameras, smartphones, and camera drones. First, the resultant *motion panorama* [3], as if virtually generated by a static camera, is by itself appealing for visual perception. More importantly, many vision tasks may benefit from GMC. For instance,

B. Leibe et al. (Eds.): ECCV 2016, Part VI, LNCS 9910, pp. 101–119, 2016.
DOI: 10.1007/978-3-319-46466-4_7

Fig. 1. Schematic diagrams of proposed TRGMC and existing sequential GMC algorithms, and resultant motion panorama for a video shot by panning the camera up and down. Background continuity breaks easily in the case of the sequential GMC [10].

dense trajectories [4] are shown to be superior when camera motion is compensated [5]. Otherwise, camera motion interferes with human motion, rendering the analysis problem very challenging. GMC allows reconstruction of a "stitched" background [6], and subsequently segmentation of foreground [7,8]. This helps multi-object tracking by mitigating the unconstrained problem of tracking multiple in-the-wild objects, to tracking objects with a static background [9].

In existing GMC works [10–12], frames are transformed to a global motion-compensated coordinate (GMCC), by *sequentially* processing input frames. For a pair of consecutive frames, the mapping transformation is estimated, and by accumulating the transformations, a *composite* global transformation of each frame to the GMCC is obtained. However, the sequential processing scheme causes frequent GMC failures for multiple reasons: (1) Sequential GMC is only as strong as the *weakest* pair of consecutive frames. A single frame with high blur or dominant foreground motion can cause the rest of the video to fail. (2) Generally, multiple planes exist in the scene. The common assumption of a single homography will accumulate residual errors into remarkable errors. (3) Even if the error of consecutive frames is in a sub-pixel scale, due to the *multiplication* of several homography matrices, the error can be significant over time [6]. These problems are especially severe when processing long videos and/or the camera motion becomes more complicated. E.g., when the camera pans to left and right repeatedly, or severe camera vibration exists, the GMC error is obvious by exhibiting discontinuity on the background (see Fig. 1 for an example).

To address the issues of sequential GMC, we propose a temporally robust global motion compensation (TRGMC) algorithm which by *joint* alignment of input frames, estimates accurate and temporally consistent transformations to GMCC. The result can be rendered as a motion panorama that maintains perceptual realism despite complicated camera motion (Fig. 1). TRGMC densely connects pairs of frames, by matching local keypoints via keypoint descriptors.

Joint alignment (a.k.a. congealing) of these frames is formulated as an optimization problem where the transformation of each frame is updated iteratively, such that for each *link* interconnecting a keypoint pair, the spatial coordinates of two end points are identical. This novel *keypoint-based congealing*, built upon succinct keypoint coordinates instead of high-dimensional appearance features, is the core of TRGMC. Joint alignment not only leads to the temporal consistency of GMC, but also improves GMC stability by using redundancy of the information. This improved stability is crucial for GMC, especially in the presence of considerable foreground motion, motion blur, non-rigid motion like water, or low-texture background. The joint alignment scheme also provides capabilities such as coarse-to-fine alignment, i.e., alignment of the keyframes followed by non-keyframes, and appropriate weighting of keypoints matches, which cannot be naturally integrated into sequential GMC. Our quantitative experiments reveal that TRGMC pushes the alignment error close to human performance.

2 Prior Work

TRGMC is related to many techniques in different aspects. We first review them and then compare our work with existing GMC algorithms.

Firstly, homography estimation from keypoint matches is crucial to many vision tasks, e.g., image stitching, registration, and GMC. Its main challenge is the false matches due to appearance ambiguities. Methods are proposed to either be robust to outliers, such as RANSAC [13–16] and reject false matches [17,18], or probabilistically combine appearance similarities and keypoint matches [10,19]. *All methods estimate a homography for a frame pair.* In contrast, we jointly estimate homographies of *all frames* to a global coordinate, which leverages the redundant background matches over time to better handle outliers.

Image stitching (IS) and panoramic image mosaicing share similarity with GMC. IS aims to minimize the distortions and ghosting artifact in the overlap region. Recent works focus on different challenges, e.g., multi-plane scenes [20–25], the parallax issue [26–28], and motion blur [29]. In these works, input images have much less overlap than GMC. On the other hand, video mosaicing takes in a video which raster scans a wide angle *static* scene, and produces a single *static* panoramic image [30–32]. When the camera path forms a 2D scan [30] or a 360° rotation [32], global refinement is performed via bundle adjustment (BA) [33], which ensures an artifact-free panoramic image. Although a byproduct of TRGMC is a similar static reconstruction of the scene, TRGMC focuses on efficient generation of an appealing video, for a *highly dynamic* scene. While one may use BA to estimate camera pose and then transformation between frames, our experiments reveal that BA is not reliable for videos with foreground motion and is less efficient than TRGMC. Hence, image/video mosaicing and GMC have different application scenarios and challenges.

Another related topic is the panoramic video [34–38]. For instance, Perazzi et al. [35] create a panoramic video from an array of stationary cameras by generalizing parallax-tolerant image stitching to video stitching. While these

works focus on stitching *multiple* synchronized videos, GMC creates a motion panorama from a *single non-stationary* camera. Unlike GMC, video panoramas do not require the resultant video to have a stationary background.

Video stabilization (VS) is a closely related but different problem. TRGMC can be re-purposed for VS, but not vice versa, due to the accuracy requirement. Given the accurate mapping to a global coordinate using TRGMC, VS would mainly amount to cropping out a smooth sequence of frames and handling rendering issues such as parallax. Among different categories of VS, 2D VS methods calculate consecutive warping between the frames and have similarities with *sequential* GMC, but any estimation error will not cause severe degradation in VS as long as it is smoothed. While TRGMC targets *long-term staticness of the background*, VS mainly cares about *smoothing* of camera motion, not *removing* it. In other words, TRGMC imposes a stronger constraint on the result. This strict requirement differentiates TRGMC also from Re-Cinematography [39].

Congealing aims to jointly align a stack of images from one object class, e.g., faces and letters [40–43]. Congealing iteratively updates the transformations of all images such that the entropy [40] or Sum of Squared Differences (SSD) [44] of the images, is minimized. However, despite many extensions of congealing [45–49], almost all prior work define the energy based on the *appearance features* of two images. Our experiments on GMC show that appearance-based congealing is inefficient and sensitive to initialization and foreground motion. Therefore, we propose a novel keypoint-based congealing algorithm minimizing the SSD of corresponding *keypoint coordinates*. Further, most prior works apply to a spatially cropped object such as faces, while we deal with complex video frames with dynamic foreground and moving background, at a higher spatial-temporal resolution. Note that [46] uses a heuristic local feature based algorithm to rigidly align object class images. In contrast we formulate the joint alignment of keypoints as an optimization problem and solve it in a principal way.

There are a few existing sequential GMC works, where the main problem is to accurately estimate a homography transformation between consecutive frames, given challenges such as appearance ambiguities, multi-plane scene, and dominant foreground [3,10,12]. Bartoli et al. [11] first estimate an approximate 4-degree-of-freedom homography, and then refine it. Sakamoto et al. [32] generate a 360° panorama from an image sequence. Assuming a 5-degree-of-freedom homography, all the homographies are optimized jointly to prevent error accumulation. In contrast, TRGMC employs an 8-degree-of-freedom homography. Although using homography in the case of considerable camera translation and large depth variation results in parallax artifacts, using a higher degrees-of-freedom homography than prior works allows TRGMC to better handle camera panning, zooming, and translation. Safdarnejad et al. [10] incorporate edge matching into a probabilistic framework that scores candidate homographies. Although [10,12] improve the robustness to foreground, error accumulation and failure in a single frame pair still deteriorate the overall performance. Thus, TRGMC targets robustness of the GMC in terms of both the presence of foreground and long-term consistency by joint alignment of frames.

3 Proposed TRGMC Algorithm

The core of TRGMC is the novel keypoint-based congealing algorithm. Our method relies on densely interconnecting the input frames, regardless of their temporal offset, by matching the detected SURF keypoints at each frame using SURF descriptors [50]. We refer to these connections, shown in Fig. 2, as *links*. Frames are initialized to their approximate spatial location by only 2D translation (Sect. 3.4). We rectify the keypoint matches such that majority of the links have end points on the background region. Then the congealing applies appropriate transformation to each frame and the links connected to it, such that the spatial coordinates of the end-points of each link are as similar as possible. In Fig. 2, this translates to having the links as parallel to the $t-$axis as possible.

For efficiency and robustness, TRGMC processes an input video in two stages. Stage one selects and jointly aligns a set of keyframes. The keyframes are frozen, and then stage two aligns each remaining frame to its two encompassing keyframes. The remainder of this section presents the details of the algorithm.

3.1 Formulation of Keypoint-Based Congealing

Given a stack of N frames $\{\mathbf{I}^{(i)}\}$, with indices $i \in \mathbb{K} = \{k_1, ..., k_N\}$, the keypoint-based congealing is formulated as an optimization problem,

$$\min_{\{\mathbf{p}_i\}} \epsilon = \sum_{i \in \mathbb{K}} [\mathbf{e}_i(\mathbf{p}_i)]^\intercal \Omega^{(i)} [\mathbf{e}_i(\mathbf{p}_i)], \tag{1}$$

where $\mathbf{p}_i$ is the transformation parameter from frame i to GMCC, $\mathbf{e}_i(\mathbf{p}_i)$ collects the pair-wise alignment errors of frame i relative to all the other frames in the stack, and $\Omega^{(i)}$ is a weight matrix.

We define the alignment error of frame i as the SSD between the spatial coordinates of the endpoints of all links connecting frame i to the other frames, instead of the SSD of appearance [44]. Specifically, as shown in Fig. 3, we denote coordinates of the start and the end point of each link k connecting frame i to the frame $d_k^{(i)} \in \mathbb{K} \backslash \{i\}$ as $(x_k^{(i)}, y_k^{(i)})$ and $(u_k^{(i)}, v_k^{(i)})$, respectively. For simplicity, we omit the frame index i in $\mathbf{p}_i$. Thus, the error $\mathbf{e}_i(\mathbf{p})$ is defined as,

$$\mathbf{e}_i(\mathbf{p}) = [\mathbf{\Delta x}_i(\mathbf{p})^\intercal, \mathbf{\Delta y}_i(\mathbf{p})^\intercal]^\intercal, \tag{2}$$

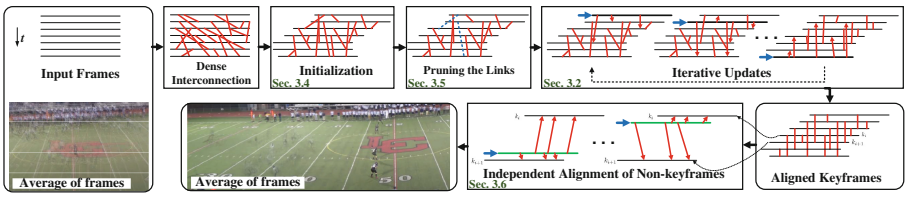

Fig. 2. Flowchart of the TRGMC algorithm.

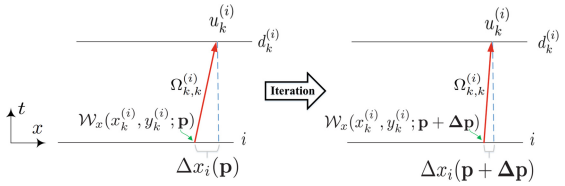

Fig. 3. The notation used in TRGMC.

where $\mathbf{\Delta x}_i(\mathbf{p}) = \mathbf{w}_x^{(i)} - \mathbf{u}^{(i)}$ and $\mathbf{\Delta y}_i(\mathbf{p}) = \mathbf{w}_y^{(i)} - \mathbf{v}^{(i)}$, are the errors in $x-$ and $y-$ axes. The vectors $\mathbf{w}_x^{(i)} = [\mathcal{W}_x(x_k^{(i)}, y_k^{(i)}; \mathbf{p})]$ and $\mathbf{w}_y^{(i)} = [\mathcal{W}_y(x_k^{(i)}, y_k^{(i)}; \mathbf{p})]$ denote the x and $y-$ coordinates of $(x_k^{(i)}, y_k^{(i)})$ warped by the parameter $\mathbf{p}$, respectively. The vectors $\mathbf{u}^{(i)} = [u_k^{(i)}]$ and $\mathbf{v}^{(i)} = [v_k^{(i)}]$ denote the coordinates of the end points and $\mathbf{x}^{(i)} = [x_k^{(i)}]$ and $\mathbf{y}^{(i)} = [y_k^{(i)}]$ denote the coordinates of the start points. If N_i links emanate from frame i, $\mathbf{e}_i$ is a $2N_i-$dim vector. $\Omega^{(i)}$ is a diagonal matrix of size $2N_i \times 2N_i$ which assigns a weight to each element in $\mathbf{e}_i$. The parameter $\mathbf{p}$ has 2, 6, or 8 elements for the cases of 2D translation, affine transformation, or homography, respectively. In this paper, we focus on homography transformation which is a projective warp model, parameterized as,

$$\begin{bmatrix} \mathcal{W}_x(x_k^{(i)}, y_k^{(i)}; \mathbf{p}) \\ \mathcal{W}_y(x_k^{(i)}, y_k^{(i)}; \mathbf{p}) \\ 1 \end{bmatrix} = \overbrace{\begin{bmatrix} p_1 & p_2 & p_3 \\ p_4 & p_5 & p_6 \\ p_7 & p_8 & 1 \end{bmatrix}}^{\mathbf{p}} \begin{bmatrix} x_k^{(i)} \\ y_k^{(i)} \\ 1 \end{bmatrix}. \tag{3}$$

Although the homography model assumes the planar scene and this assumption may be violated in real world [27], we identify the problem of temporal robustness to be more fundamental for GMC than the inaccuracies due to a *single* homography. Also, videos for GMC are generally swiped through the scene with high overlap, thus the discontinuity resulted from this assumption is minor.

3.2 Optimization solution

Equation 1 is a non-linear optimization problem and difficult to minimize. Following [44], we linearize this equation by taking the first-order Taylor expansion around $\mathbf{p}$. Starting from an initial $\mathbf{p}$, the goal is to estimate $\Delta\mathbf{p}$ by,

$$\underset{\Delta\mathbf{p}}{\operatorname{argmin}} [\mathbf{e}_i(\mathbf{p}) + \frac{\partial \mathbf{e}_i(\mathbf{p})}{\partial \mathbf{p}} \Delta\mathbf{p}]^\intercal \Omega^{(i)} [\mathbf{e}_i(\mathbf{p}) + \frac{\partial \mathbf{e}_i(\mathbf{p})}{\partial \mathbf{p}} \Delta\mathbf{p}] + \gamma \Delta\mathbf{p}^\intercal \mathcal{I} \Delta\mathbf{p}, \tag{4}$$

where $\Delta\mathbf{p}^\intercal \mathcal{I} \Delta\mathbf{p}$ is a regularization term, with a positive constant γ setting the trade-off. We observe that without this regularization, parameter estimation may lead to distortion of the frames. The indicator matrix $\mathcal{I}$ is a diagonal matrix specifying which elements of $\Delta\mathbf{p}$ need a constraint. We use

$\mathcal{I} = diag([1,1,0,1,1,0,1,1])$ to specify that there is no constraint on the translation parameters of the homography, but the rest of parameters should remain small.

By setting the first-order derivative of Eq. 4 to zero, the solution for $\Delta\mathbf{p}$ is,

$$\Delta\mathbf{p} = \mathbf{H}_R^{-1} \frac{\partial \mathbf{e}_i(\mathbf{p})^\intercal}{\partial \mathbf{p}} \Omega^{(i)} \mathbf{e}_i(\mathbf{p}), \tag{5}$$

$$\mathbf{H}_R = \frac{\partial \mathbf{e}_i(\mathbf{p})^\intercal}{\partial \mathbf{p}} \Omega^{(i)} \frac{\partial \mathbf{e}_i(\mathbf{p})}{\partial \mathbf{p}} + \gamma\mathcal{I}. \tag{6}$$

Using the chain rule, we have $\frac{\partial \mathbf{e}_i(\mathbf{p})}{\partial \mathbf{p}} = \frac{\partial \mathbf{e}_i(\mathbf{p})}{\partial \mathcal{W}} \frac{\partial \mathcal{W}}{\partial \mathbf{p}}$. Knowing that the mapping has two components as $\mathcal{W} = (\mathcal{W}_x, \mathcal{W}_y)$, and the first half of $\mathbf{e}_i$ only contains x components and the rest only y components, we have,

$$\frac{\partial \mathbf{e}_i(\mathbf{p})}{\partial \mathcal{W}} = \begin{bmatrix} \mathbf{1}_{N_i} & \mathbf{0}_{N_i} \\ \mathbf{0}_{N_i} & \mathbf{1}_{N_i} \end{bmatrix}, \tag{7}$$

where $\mathbf{1}_{N_i}$ (or $\mathbf{0}_{N_i}$) is a N_i-dim vector with all elements being 1 (or 0). For homography transformation, $\frac{\partial \mathcal{W}}{\partial \mathbf{p}} = \frac{\partial(\mathcal{W}_x, \mathcal{W}_y)}{\partial(p_1,p_2,p_3,p_4,p_5,p_6,p_7,p_8)}$ is given by,

$$\frac{\partial \mathcal{W}}{\partial \mathbf{p}} = \begin{bmatrix} \mathbf{w}_x^{(i)} & \mathbf{w}_y^{(i)} & \mathbf{1}_{N_i} & \mathbf{0}_{N_i} & \mathbf{0}_{N_i} & \mathbf{0}_{N_i} & -\mathbf{u}^{(i)}\mathbf{w}_x^{(i)} & -\mathbf{u}^{(i)}\mathbf{w}_y^{(i)} \\ \mathbf{0}_{N_i} & \mathbf{0}_{N_i} & \mathbf{0}_{N_i} & \mathbf{w}_x^{(i)} & \mathbf{w}_y^{(i)} & \mathbf{1}_{N_i} & -\mathbf{v}^{(i)}\mathbf{w}_x^{(i)} & -\mathbf{v}^{(i)}\mathbf{w}_y^{(i)} \end{bmatrix}. \tag{8}$$

At each iteration, and for each frame i, $\Delta\mathbf{p}$ is calculated and the start points of all the links emanating from frame i are updated accordingly. Similarly, for all links with end points on frame i, the end point coordinates are updated.[1]

We use the SURF [50] algorithm for keypoint detection with a low detection threshold, $\tau_s = 200$, to ensure sufficient keypoints are detected even for low-texture backgrounds. We use the nearest-neighbor ratio method [51] to match the keypoints descriptors and form links between each pair of keyframes.

Keyframe selection. We select keyframes at a constant step of Δf, i.e., from every Δf frames, only one is selected. Based on the experimental results, as a trade-off between accuracy and efficiency, we use $\Delta f = 10$ in TRGMC.

3.3 Weight assignment

We have defined all parameters in the problem formulation, except the weights of links, $\Omega^{(i)}$. We consider two factors in setting $\Omega^{(i)}$. Firstly, the keypoints detected at larger scales are more likely to be from background matches, since they cover coarser information and larger image patches. Thus, to be robust to foreground,

[1] In algorithm implementation, it is important to store the original coordinates of the detected keypoints and apply the *composite* transformations accumulated in all the iterations to update the coordinates of the start and end points of the links. Otherwise, accumulation of numerical errors will harm the performance.

the early iterations should emphasize links from larger-scale keypoints, which forms a coarse-to-fine alignment. We normalize the scales of all keypoints such that the maximum is 1, and denote the minimum of the normalized scales of the two keypoints comprising the link k as s_k. Then, $\Omega_{k,k}^{(i)}$ is set proportional to s_k.

Secondly, for each frame i, the links may be made either to all the previous frames, denoted as *backward* scheme, or both the previous and upcoming frames, denoted as *backward-forward* scheme. The former is for real-time applications, whereas the latter for offline video processing. These schemes are implemented by assigning different weights to backward and forward links,

$$\Omega_{k,k}^{(i)} = \begin{cases} (\beta.s_k)^{r^q}; \text{ if } d_k^{(i)} < i & \text{(Backward links)} \\ (\alpha.s_k)^{r^q}; \text{ if } d_k^{(i)} > i & \text{(Forward links)} \end{cases} \tag{9}$$

where $0 < \alpha, \beta < 1$, q is the iteration index, and $0 < r < 1$ is the rate of change of the weights. Note that the alignment errors in x and $y-$axes have the same weights, i.e., $\Omega_{k+N_i,k+N_i}^{(i)} = \Omega_{k,k}^{(i)}$. After a few iterations, the weights of all the links will be restored to 1. In the backward scheme, we set $\alpha = 0$.

3.4 Initialization

Initialization speeds up the alignment and decreases the false keypoint matches. The objective is to roughly place each frame at the appropriate coordinates in the GMCC. For initialization, we align the frames based only on rough estimation of translation without considering rotation, skew, or scale. We use the average of the motion vectors in matching two consecutive frames as the translation. Using this simple initialization, even if the camera has in-plane rotation, estimated 2D translations are zero, which is indeed correct and does not cause any problem for TRGMC. Given the estimated translation, approximate overlap area of each pair of frames is calculated, and only the keypoints inside the overlap area are matched, reducing number of false matches due to appearance ambiguities.

3.5 Outlier handling

Links may become outliers for two reasons: (i) the keypoints reside on foreground objects not consistent with camera motion; (ii) false links between different physical locations are caused by the low detection threshold and similar appearances.

In order to prune the outliers, we assume that the motion vectors of background matches, i.e., background links, have consistent and smooth patterns, caused by camera motion such as pan, zoom, tilt, whereas, the outlier links will exhibit arbitrary pattern, inconsistent with the background pattern. Specifically, we use Ma et al. [17] method to prune outlier links by imposing a smoothness constraint on the motion vector field[2]. This method outperforms RANSAC if the set of keypoint matches contains a large proportion of outliers. Since keyframes have larger relative time difference than consecutive frames, the foreground motion

[2] We use the implementation provided by the authors and the default parameters.

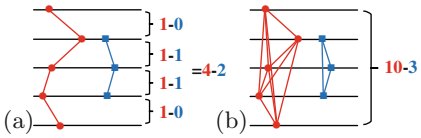

Fig. 4. Comparison of the ratios of background-foreground matches for (a) sequential GMC and (b) TRGMC.

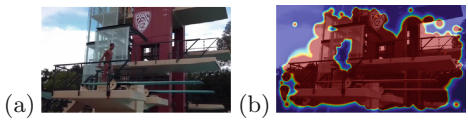

Fig. 5. (a) The input frame, (b) the reliability map, with the red color showing higher reliability. (Color figure online)

is accentuated and more distinguishable from camera motion. This helps with better pruning of the foreground links. At each stage that the keypoints from a pair of frames are matched to form the links, we perform the pruning.

Congealing of an image stack also increases the proportion of background matches over the outliers - another way to suppress outliers. The keypoints on background are more likely to form longer range matches than the foreground ones, due to non-rigid foreground motion. Hence, when $\binom{N}{2}$ combinatorial pairs of frames are interconnected, there are a lot more background matches (Fig. 4).

3.6 Alignment of Non-keyframes

The keyframes alignment provides a set of temporally consistent motion compensated frames, which are the basis for aligning non-keyframes. We refer to keyframes and non-keyframes with superscripts i and j, respectively. For a non-keyframe j between the keyframes k_i and k_{i+1}, its alignment is a special case of Eq. 1, with indices $\mathbb{K} = \{j\}$, and the destination of the links $d_k^{(j)} \in \{k_i, k_{i+1}\}$, i.e., only $\mathbf{p}_j$ of frame j is updated while the keyframes remain fixed. Each non-keyframe between keyframes k_i and k_{i+1} is aligned independently.

However, given the small time offset between j and $d_k^{(j)}$, the observed foreground motion may be hard to discern. Also, frame j is linked only to two keyframes, thus there is no redundancy of background information to improve robustness to foreground motion. So, we handle outlier handling by assigning higher weights to links that are more likely to be connected to the background.

For each keyframe i, we quantify how well the links emanating from frame i are aligned with other keyframes. If the alignment error is small, i.e., $\epsilon_k^{(i)} = \left|\mathcal{W}_x(x_k^{(i)}, y_k^{(i)}; \mathbf{p}) - u_k^{(i)}\right| + \left|\mathcal{W}_y(x_k^{(i)}, y_k^{(i)}; \mathbf{p}) - v_k^{(i)}\right| < \tau$, the link k is more likely on the background of frame i and thus, more reliable for aligning non-keyframes. We create a *reliability map* for each keyframe i, denoted as $\mathbf{R}^{(i)}$ (Fig. 5). For each link k with $\epsilon_k^{(i)} < \tau$, a Gaussian function with $\mu_k = (x_k^{(i)}, y_k^{(i)})$ and $\sigma_k = cs_k$ is superposed on $\mathbf{R}^{(i)}$, where the constant c is 20. We define,

$$\mathbf{R}_{m,n}^{(i)} = \left\lceil \left\lfloor \sum_{k \in \mathbb{B}_i} e^{-\frac{\left(m - x_k^{(i)}\right)^2 + \left(n - y_k^{(i)}\right)^2}{2\sigma_k^2}} \right\rfloor_1 \right\rceil_\eta , \tag{10}$$

where $\mathbb{B}_i = \{k|\epsilon_k^{(i)} < \tau\}$, $\eta > 0$ is a small constant (set to 0.1), $\lceil x \rceil_\eta = \max(x, \eta)$ and $\lfloor x \rfloor_1 = \min(1, \eta)$. Now, we assign the weight of the links connecting frame j to the keyframe $d_k^{(j)}$ at the coordinate $(u_k^{(j)}, v_k^{(j)})$, as the reliability map of the keyframe at the endpoint, $\Omega_{k,k}^{(j)} = \left(\mathbf{R}^{(a)}_{u_k^{(j)}, v_k^{(j)}}\right)^{r^q}$, where $a = d_k^{(j)}$.

We summarize the TRGMC algorithm in Algorithm 1.

Algorithm 1. TRGMC Algorithm

Data: A set of input frames $\{\mathbf{I}^{(m)}\}_{m=1}^{M}$
Result: A set of homography matrices $\{\mathbf{p}_m\}_{m=1}^{M}$
`/* Align keyframes (Sec. 3.2 )` `*/`
Specify $\mathbb{K} = \{k_1, ..., k_N\}$ and initialize (Sec. 3.4);
Match keypoints of all frames $i \in \mathbb{K}$ densely;
Prune links (Sec. 3.5) and set weights (Eqn. 9);
Store links' start and end coordinates in $(\mathbf{x}_i, \mathbf{y}_i)$ and $(\mathbf{u}_i, \mathbf{v}_i)$;
repeat
  forall $i \in \mathbb{K}$ **do**
    Compute $\Delta\mathbf{p}_i$ (Eqn. 5), update $\mathbf{p}_i$, $\mathbf{x}_i$ and $\mathbf{y}_i$;
    Update $(\mathbf{u}_m, \mathbf{v}_m)$ according to $\mathbf{p}_i$ for $m \in \mathbb{K} \backslash \{i\}$;
    Update weights (Eqn. 9);
  $q \leftarrow q + 1$;
until $q < T_1$ *or* $\left(\frac{1}{N}\sum_{i\in\mathbb{K}} ||\Delta\mathbf{p}_i||^2 > \tau_1\right)$;
`/* Align non-keyframes (Sec. 3.6)` `*/`
Compute reliability map $\mathbf{R}^{(i)}$ for $i \in \mathbb{K}$;
for $i = 1 : N - 1$ **do**
  forall $j \in \{k_i + 1, ..., k_{i+1} - 1\}$ **do**
    Match keypoints in j with $d^{(j)} \in \{k_i, k_{i+1}\}$;
    Prune links (Sec. 3.5) and set weights $\Omega_{k,k}^{(j)}$;
    Store links' coordinates in $(\mathbf{x}_j, \mathbf{y}_j)$ and $(\mathbf{u}_j, \mathbf{v}_j)$;
    repeat
      Compute $\Delta\mathbf{p}_j$ (Eqn. 5), update $\mathbf{p}_j$, $\mathbf{x}_j$ and $\mathbf{y}_j$;
      Update weights (Eqn. 9), $q \leftarrow q + 1$;
    until $q < T_2$ *or* $\left(||\Delta\mathbf{p}_j||^2 > \tau_2\right)$;

4 Experimental Results and Applications

We now present qualitative and quantitative results of the TRGMC algorithm and discuss how different computer vision applications will benefit from TRGMC.

4.1 Experiments and results

Baselines and details. We choose three sequential GMC algorithms as the baselines for comparison: MLESAC [15] and HEASK [19] both based on our own implementation, and RGMC [10] based on the authors' Matlab code available online. TRGMC is implemented in Matlab and is available for download.[3] Denoting the video frames of $w \times h$ pixels, we set the parameters as $\gamma = 0.1wh$, $T_1 = 300$, $\tau_1 = 5 \times 10^{-4}$, $T_2 = 50$, $\tau_2 = 10^{-4}$, $r = 0.7$, $\tau = 1$, $\Delta f = 10$, and $\beta = 1$. For the backward-forward scheme we set $\alpha = 1$ and for the backward scheme $\alpha = 0$.

Datasets and metric. We form a dataset composed of 40 challenging videos from SVW [52] and 15 videos from UCF101 [53], termed "quantitative dataset". SVW is an extremely unconstrained dataset including videos of amateurs practicing sports, and is also captured by amateurs via smartphone. In addition, we form another "qualitative dataset" with 200 *unlabeled* videos from SVW, in challenging categories of boxing, diving, and hockey.

To compare GMC over different temporal distances of frames, for each video of length M frames in the quantitative dataset, we manually align all 10 possible pairs from the 5-frame set, $\mathbb{F} = \{1, 0.25M, 0.5M, 0.75M, M\}$, as long as they are overlapping, and specify the background regions. For this, a GUI is developed for a labeler to match 4 points on each frame pair, and fine tune them up to a half-pixel accuracy, until the background difference is minimized. Then, the labeler selects the foreground regions which subsequently identify the background region. Similar to [10], we quantify the consistency of two warped frames $\mathbf{I}^{(i)}(\mathbf{p}_i)$ and $\mathbf{I}^{(j)}(\mathbf{p}_j)$ (0 to 1 grayscale pixels) via the background region error (BRE),

$$\mathrm{BRE}(i,j) = \frac{1}{\|\mathbf{M_B}\|_1} \| |(\mathbf{I}^{(i)}(\mathbf{p}_i) - \mathbf{I}^{(j)}(\mathbf{p}_j))| \odot \mathbf{M_B} \|_1, \tag{11}$$

where $\odot$ is element-wise multiplication and $\mathbf{M_B}$ is the background mask for the intersection of two warped frames.

Quantitative evaluation. Average of BRE over all the temporal frames pairs is shown in Table 1. TRGMC outperform all the baseline methods with considerable margin. The *backward-forward (BF)* scheme has a slightly better accuracy

Table 1. Comparison of GMC algorithms on quantitative dataset (*GT: Ground truth, BF: Backward-Forward, B: Backward).

Algorithm	MLESAC	HEASK	RGMC	TRGMC		GT*
Setting	–	–	–	BF*	B*	–
Avg. BRE	0.116	0.110	0.097	**0.058**	0.060	0.038
Efficiency (s/f)	0.17	7.47	3.47	0.64	0.41	–

[3] http://cvlab.cse.msu.edu/project-trgmc.html.

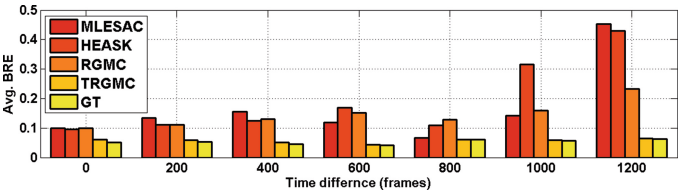

Fig. 6. Average BRE of frame pairs versus the time difference between the two frames.

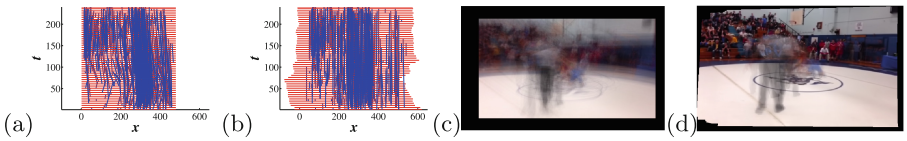

Fig. 7. Top view of the frames and links (a) before and (b) after TRGMC. The parallel links in (b) show successful *spatial* alignment of keypoints. For better visibility, we show up to 15 links emanated per frame. Average of frames (c) before and (d) after TRGMC.

Fig. 8. Composite image formed by overlaying the frame n on frame 1 for several videos after TRGMC. Left to right, n is equal to 144, 489, 912, 93, respectively. In the overlap region the difference between the frames is shown.

than the *backward (B)* scheme, and is also more stable based on our visual observation. Thus, we use BF as the default scheme for TRGMC.

To illustrate how the accumulation of errors over time affects the final error, Fig. 6 summarizes the average error versus the time difference between the frames in $\mathbb{F}$. This shows that TRGMC error is almost constant over a wide temporal distance between the frames. Thus, even if a frame is not aligned accurately, the error is not propagating to all the frames after that. However, in sequential GMC, the error increases as the time difference increases.

Qualitative evaluation. While quantitative results are comprehensive, the number of videos is limited by the labeling cost. Thus, we further compare TRGMC and the best performing baseline, RGMC, on the larger qualitative dataset. The resultant motion panoramas were *visually* investigated and categorized into three cases: good, shaking, and failed (i.e., considerable background discontinuity). The comparison in Table 2 again shows the superiority of TRGMC.

Figure 7 shows the *links* of a sample video processed by TRGMC, and the average frames, before and after processing. Initialization module is disable for generating this figure to better illustrate how well the spatial coordinate of the

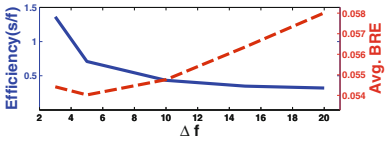

Fig. 9. Error and efficiency vs. the keyframe selection step, Δf.

Table 2. Comparison of GMC algorithms on qualitative dataset.

Alg.\Perfromance	Good	Shaking	Failed
RGMC	64 %	33 %	3 %
TRGMC	**93 %**	5 %	2 %

keypoints are aligned, resulting in links parallel to the $t-$ axis. Figure 8 shows a composite image formed by overlaying the last frame (or a far apart frame with enough overlap) on frame 1 for several videos, after TRGMC. In the overlap region, difference between the two frames is shown, to demonstrate how well the background region matches for the frames with large temporal distances.

Computational efficiency. Table 1 also presents the average time for processing one frame for each method, on a PC with an Intel i5-3470@3.2 GHz CPU, and 8 GB RAM. While obtaining considerably better accuracy than HEASK or RGMC, TRGMC is on average 15 times faster than HEASK and 7 times faster than RGMC. MLESAC is ∼3 times faster than TRGMC, but with twice the error. For TRGMC, the backward scheme is 50 % faster than forward-backward, since it has approximately half the links of BF.

Accuracy vs. efficiency trade-off. Figure 9 presents the error and efficiency results for a set of 5 videos versus the keyframe selection step, Δf. For this set, the ground truth error is 0.049. As a sweet spot in the error and efficiency trade-off, we use $\Delta f = 10$ for TRGMC. This figure also justifies the two stage processing scheme in TRGMC, as processing frames at a low selection step Δf, is costly in terms of efficiency, but only improves the accuracy slightly.

4.2 TRGMC applications

Motion panorama. By sequentially reading input frames, applying the transformation found by TRGMC, and overlaying the warped frames on a sufficiently large canvas, a motion panorama is generated. Furthermore, it is possible to reconstruct the background using the warped frames *first* (as will be discussed later), and overlay the frames on that, to create a more impressive panorama. Figure 10 shows a few exemplar panoramas and the camera motion trajectory.

Background reconstruction. Background reconstruction is important for removing occlusions, or detecting foreground [6]. To reconstruct the background, a weighted average scheme is used to weight each frame by the *reliability map*, $\mathbf{R}^{(i)}$, which assigns higher weights to background. Since the minimum value of $\mathbf{R}^{(i)}$ is a positive constant η, if no reliable keyframe exists at a coordinate, all the frames will have equal weights. Specifically, the background is reconstructed by $\mathbf{B} = \frac{\sum_{i\in\mathbb{K}} \mathbf{R}^{(i)}(\mathbf{p}_i)\mathbf{I}^{(i)}(\mathbf{p}_i)}{\sum_{i\in\mathbb{K}} \mathbf{R}^{(i)}(\mathbf{p}_i)}$, where $\mathbf{R}^{(i)}(\mathbf{p}_i)$ and $\mathbf{I}^{i}(\mathbf{p}_i)$ are the reliability map and

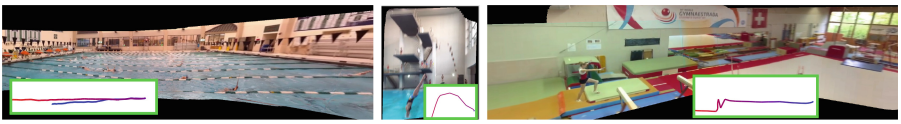

Fig. 10. Temporal overlay of frames from different videos processed by TRGMC. Trajectory of the center of image plane over time is overlaid on each plot to show the camera motion pattern, where color changes from blue to red with progression of time. (Color figure online)

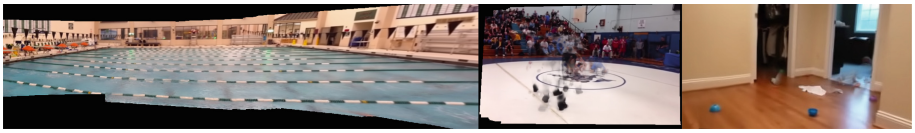

Fig. 11. Background reconstruction results. Compare the left image with Fig. 10, middle image with Fig. 7, and right image with Fig. 8.

Fig. 12. Segmented foreground overlaid on the input.

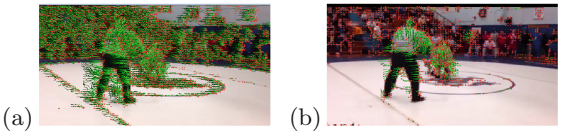

Fig. 13. Dense trajectories of the (a) original video, and (b) TRGMC-processed video.

the input frame warped using the transformation $\mathbf{p}_i$. Using our scheme, reconstructed background in Fig. 11 is sharper and less impacted by the foreground.

Foreground segmentation. The reliable background reconstruction result $\mathbf{B}$ along with the GMC result of frame $\mathbf{I}^{(i)}$, e.g., $\mathbf{p}_i$, can be easily used to segment the foreground by thresholding the difference, $|\mathbf{B} - \mathbf{I}^{(i)}(\mathbf{p}_i)|$ (Fig. 12).

Human action recognition. State of the art human action recognition heavily relies on analysis of human motion. GMC helps to suppress camera motion and magnify human motion, making the motion analysis more feasible, which is clearly shown by the dense trajectories [4] in Fig. 13.

Multi-object Tracking (MOT). When appearance cues for tracking are ambiguous, e.g., tracking players in team sports like football, motion cues gain extra significance [54,55]. MOT is comprised of two tasks, data association by assigning each detection a label, and trajectory estimation – both highly affected by camera motion. TRGMC can be applied to remove camera motion and thus, revive the power of tracking algorithms relying on motion cues. To verify the impact of TRGMC, we manually label the locations of all players in 566 frames of a football video and use this ground truth detection results to study how MOT using [56] benefits from TRGMC. Figure 14 compares the trajectories of players over time with and without applying TRGMC. Comparing number of

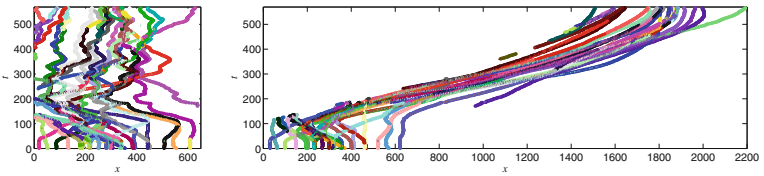

Fig. 14. Multi-player tracking using [56] for a football video with camera panning to the right, before (left) and after processing by TRGMC (right).

label switches qualitatively demonstrates improvement of a challenging MOT scenario using TRGMC. Also, the Multi-Object Tracking Accuracy [57] for the original video and the video processed by TRGMC are 63.79 % and 84.23 %, respectively.

5 Conclusions and Discussions

We proposed a temporally robust global motion compensation (TRGMC) algorithm by joint alignment (congealing) of frames, in contrast to the common sequential scheme. Despite complicated camera motions, TRGMC can remove the *intentional* camera motion, such as pan, as well as *unwanted* motion due to vibration on handheld cameras. Experiments demonstrate that TRGMC outperforms existing GMC methods, and applications of TRGMC.

The enabling assumption of TRGMC is that the camera motion in the direction of the optical axis is negligible. For instance, TRGMC will not work properly on a video from a wearable camera of a pedestrian, since in the global coordinate the upcoming frames grow in size and cause computational and rendering problems. The best results are achieved if the optical center of the camera has negligible movement, making a homography-based approximation of camera motion appropriate. However, if the optical center moves in the perpendicular direction to the optical axis (e.g., a camera following a swimmer), TRGMC still works well, but results will be visually degraded by the parallax effect.

Acknowledgement. This work was partially supported by TechSmith Corporation.

References

1. He, Y., Feng, B., Yang, S., Zhong, Y.: Fast global motion estimation for global motion compensation coding. In: Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), vol. 2, pp. 233–236. IEEE (2001)
2. Smolić, A., Vatis, Y., Schwarz, H., Wiegand, T.: Long-term global motion compensation for advanced video coding. In: ITG-Fachtagung Dortmunder Fernsehseminar, pp. 213–216 (2003)
3. Bartoli, A., Dalal, N., Bose, B., Horaud, R.: From video sequences to motion panoramas. In: Proceedings of the Conference Motion and Video Computing Workshops, pp. 201–207. IEEE (2002)

4. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176. IEEE (2011)
5. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 3551–3558. IEEE (2013)
6. Monari, E., Pollok, T.: A real-time image-to-panorama registration approach for background subtraction using pan-tilt-cameras. In: Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 237–242. IEEE (2011)
7. Sun, Y., Li, B., Yuan, B., Miao, Z., Wan, C.: Better foreground segmentation for static cameras via new energy form and dynamic graph-cut. In: Proceedings of the International Conference on Pattern Recognition (ICPR), vol. 4, pp. 49–52. IEEE (2006)
8. Wan, C., Yuan, B., Miao, Z.: A new algorithm for static camera foreground segmentation via active coutours and GMM. In: Proceedings of the International Conference on Pattern Recognition (ICPR), pp. 1–4. IEEE (2008)
9. Solera, F., Calderara, S., Cucchiara, R.: Learning to divide and conquer for online multi-target tracking. arXiv preprint arXiv:1509.03956 (2015)
10. Safdarnejad, S.M., Liu, X., Udpa, L.: Robust global motion compensation in presence of predominant foreground. In: Proceedings of the British Machine Vision Conference (BMVC) (2015)
11. Bartoli, A., Dalal, N., Horaud, R.: Motion panoramas. Comput. Animation Virtual Worlds **15**(5), 501–517 (2004)
12. Déniz, O., Bueno, G., Bermejo, E., Sukthankar, R.: Fast and accurate global motion compensation. Pattern Recogn. **44**(12), 2887–2901 (2011)
13. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. ACM Commun. **24**(6), 381–395 (1981)
14. Chum, O., Matas, J., Kittler, J.: Locally optimized RANSAC. In: Michaelis, B., Krell, G. (eds.) DAGM 2003. LNCS, vol. 2781, pp. 236–243. Springer, Heidelberg (2003). doi:10.1007/978-3-540-45243-0_31
15. Torr, P.H., Zisserman, A.: MLESAC: a new robust estimator with application to estimating image geometry. Comput. Vis. Image Underst. **78**(1), 138–156 (2000)
16. Tordoff, B.J., Murray, D.W.: Guided-MLESAC: faster image transform estimation by using matching priors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1523–1535 (2005)
17. Ma, J., Zhao, J., Tian, J., Yuille, A.L., Tu, Z.: Robust point matching via vector field consensus. IEEE Trans. Image Process. **23**(4), 1706–1721 (2014)
18. Li, X., Hu, Z.: Rejecting mismatches by correspondence function. Int. J. Comput. Vis. **89**(1), 1–17 (2010)
19. Yan, Q., Xu, Y., Yang, X., Nguyen, T.: HEASK: Robust homography estimation based on appearance similarity and keypoint correspondences. Pattern Recogn. **47**(1), 368–387 (2014)
20. Szpak, Z.L., Chojnacki, W., van den Hengel, A.: Robust multiple homography estimation: An ill-solved problem. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2132–2141. IEEE (2015)
21. Zuliani, M., Kenney, C.S., Manjunath, B.: The multiRANSAC algorithm and its application to detect planar homographies. In: Proceedings of the International Conference on Image Processing (ICIP), vol. 3, pp. III–153. IEEE (2005)

22. Toldo, R., Fusiello, A.: Robust multiple structures estimation with J-Linkage. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 537–547. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88682-2_41
23. Ma, J., Chen, J., Ming, D., Tian, J.: A mixture model for robust point matching under multi-layer motion. PloS One **9**(3), e92282 (2014)
24. Uemura, H., Ishikawa, S., Mikolajczyk, K.: Feature tracking and motion compensation for action recognition. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 1–10 (2008)
25. Gao, J., Kim, S.J., Brown, M.S.: Constructing image panoramas using dual-homography warping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 49–56. IEEE (2011)
26. Lin, W.Y., Liu, S., Matsushita, Y., Ng, T.T., Cheong, L.F.: Smoothly varying affine stitching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 345–352. IEEE (2011)
27. Zaragoza, J., Chin, T.J., Tran, Q.H., Brown, M.S., Suter, D.: As-projective-as-possible image stitching with moving dlt. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1285–1298 (2014)
28. Lin, C.C., Pankanti, S.U., Ramamurthy, K.N., Aravkin, A.Y.: Adaptive as-natural-as-possible image stitching. In: Proceedings of the IEEE Conference Computer Vision and Pattern Recognition (CVPR), pp. 1155–1163. IEEE (2015)
29. Li, Y., Kang, S.B., Joshi, N., Seitz, S.M., Huttenlocher, D.P.: Generating sharp panoramas from motion-blurred videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2424–2431. IEEE (2010)
30. Sawhney, H.S., Hsu, S., Kumar, R.: Robust video mosaicing through topology inference and local to global alignment. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 103–119. Springer, Heidelberg (1998). doi:10.1007/BFb0054736
31. Shum, H.Y., Szeliski, R.: Construction and refinement of panoramic mosaics with global and local alignment. In: Proceedings of the International Conference on Computer Vision (ICCV)
32. Sakamoto, M., Sugaya, Y., Kanatani, K.: Homography optimization for consistent circular panorama generation. In: Chang, L.-W., Lie, W.-N. (eds.) PSIVT 2006. LNCS, vol. 4319, pp. 1195–1205. Springer, Heidelberg (2006). doi:10.1007/11949534_121
33. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment — a modern synthesis. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) IWVA 1999. LNCS, vol. 1883, pp. 298–372. Springer, Heidelberg (2000). doi:10.1007/3-540-44480-7_21
34. El-Saban, M., Izz, M., Kaheel, A., Refaat, M.: Improved optimal seam selection blending for fast video stitching of videos captured from freely moving devices. In: Proceedings of the International Conference on Image Processing (ICIP), pp. 1481–1484. IEEE (2011)
35. Perazzi, F., Sorkine-Hornung, A., Zimmer, H., Kaufmann, P., Wang, O., Watson, S., Gross, M.: Panoramic video from unstructured camera arrays. In: Computer Graphics Forum, vol. 34, pp. 57–68. Wiley Online Library (2015)
36. Zeng, W., Zhang, H.: Depth adaptive video stitching. In: Proceedings of the IEEE Conference on Computer and Information Science (ICIS), pp. 1100–1105. IEEE (2009)
37. Jiang, W., Gu, J.: Video stitching with spatial-temporal content-preserving warping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 42–48. IEEE (2015)

38. Ibrahim, M.T., Hafiz, R., Khan, M.M., Cho, Y., Cha, J.: Automatic reference selection for parametric color correction schemes for panoramic video stitching. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Fowlkes, C., Wang, S., Choi, M.-H., Mantler, S., Schulze, J., Acevedo, D., Mueller, K., Papka, M. (eds.) ISVC 2012. LNCS, vol. 7431, pp. 492–501. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33179-4_47
39. Gleicher, M.L., Liu, F.: Re-cinematography: Improving the camerawork of casual video. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **5**(1), 2 (2008)
40. Learned-Miller, E.G.: Data driven image models through continuous joint alignment. IEEE Trans. Pattern Anal. Mach. Intell. **28**(2), 236–250 (2006)
41. Liu, X., Tong, Y., Wheeler, F.W.: Simultaneous alignment and clustering for an image ensemble. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 1327–1334. IEEE (2009)
42. Tong, Y., Liu, X., Wheeler, F.W., Tu, P.: Automatic facial landmark labeling with minimal supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2009)
43. Liu, X., Tong, Y., Wheeler, F.W., Tu, P.H.: Facial contour labeling via congealing. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 354–368. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15549-9_26
44. Cox, M., Sridharan, S., Lucey, S., Cohn, J.: Least squares congealing for unsupervised alignment of images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8. IEEE (2008)
45. Huang, G., Mattar, M., Lee, H., Learned-Miller, E.G.: Learning to align from scratch. In: Advances in Neural Information Processing Systems (NIPS), pp. 764–772 (2012)
46. Lankinen, J., Kämäräinen, J.K.: Local feature based unsupervised alignment of object class images. In: Proceedings of the British Machine Vision Conference (BMVC), vol. 1 (2011)
47. Lucey, S., Navarathna, R., Ashraf, A.B., Sridharan, S.: Fourier Lucas-Kanade algorithm. IEEE Trans. Pattern Anal. Mach. Intell. **35**(6), 1383–1396 (2013)
48. Cox, M., Sridharan, S., Lucey, S., Cohn, J.: Least-squares congealing for large numbers of images. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 1949–1956. IEEE (2009)
49. Shokrollahi Yancheshmeh, F., Chen, K., Kamarainen, J.K.: Unsupervised visual alignment with similarity graphs. In: Proceedings of the IEEE Conference Computer Vision and Pattern Recognition (CVPR), pp. 2901–2908. IEEE (2015)
50. Bay, H., Tuytelaars, T., Gool, L.: SURF: speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006). doi:10.1007/11744023_32
51. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
52. Safdarnejad, S.M., Liu, X., Udpa, L., Andrus, B., Wood, J., Craven, D.: Sports videos in the wild (SVW): a video dataset for sports analysis. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–7. IEEE (2015)
53. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
54. Lezama, J., Alahari, K., Sivic, J., Laptev, I.: Track to the future: spatio-temporal video segmentation with long-range motion cues. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2011)

55. Dicle, C., Camps, O., Sznaier, M.: The way they move: tracking multiple targets with similar appearance. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 2304–2311. IEEE (2013)
56. Andriyenko, A., Schindler, K., Roth, S.: Discrete-continuous optimization for multi-target tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1926–1933. IEEE (2012)
57. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the CLEAR MOT metrics. J. Image Video Process. **2008**, 1 (2008)