Video Synchronization Based on Projective-Invariant Descriptor

Qiang Zhang^{1,2} · Lin Yao² · Yajun Li² · Jungong Han³

Published online: 25 July 2018
© The Author(s) 2018

Abstract

In this paper, we present a novel trajectory-based method to synchronize two videos shooting the same dynamic scene, which are recorded by stationary un-calibrated cameras from different viewpoints. The core algorithm is carried out in two steps: projective-invariant descriptor construction and trajectory points matching. In the first step, a new five-coplanar-points structure is proposed to compute the cross ratio during the construction of the projective-invariant descriptor. The five points include one trajectory point and four fixed points induced from the background scene, which are *co-planar* in the 3D coordinate. In the second step, the matched trajectory points are initially estimated by the primitive nearest neighbor method, and are further refined by using epipolar geometric constraints and post processing. Experimental results demonstrate that the proposed method significantly outperforms the existing state-of-the-arts. More importantly, the proposed method is more generic in the sense that it works well for those videos captured under different conditions, including different frame rates, wide baseline, multiple moving objects, planar or non-planar motion trajectories.

Keywords Video synchronization · Projective-invariant descriptor · Cross ratio · Five-coplanar-points structure · Non-planar motion

1 Introduction

Video synchronization is a technique that temporally aligns two videos, which are recorded for the same dynamic scene by un-calibrated cameras from distinct viewpoints or over different time intervals [1]. The critical goal for video synchronization is to establish temporal correspondences among frames of two input videos, i.e., a reference video and a video to be synchronized. The applications of video synchronization cover a wide range of video anal-

✉ Jungong Han
jungong.han@lancaster.ac.uk

¹ Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an 710071, Shaanxi, China

² Center for Complex Systems, School of Mechano-electronic Engineering, Xidian University, Xi'an 710071, Shaanxi, China

³ InfoLab21, School of Computing and Communications, Lancaster University, Lancaster LA1 4YW, UK

ysis tasks [2–8], such as video surveillance, target identification, human action recognition, saliency detection and fusion.

For simplicity, the temporal alignment is assumed to be modeled by a 1D affine transformation¹ in this paper. The relationship between the frame index i in the reference video and its corresponding frame index j in the video to be synchronized is assumed to satisfy:

$$j = \alpha \times i + \Delta t \quad (1)$$

where α is the ratio of the frame rates between the two videos, and Δt is the temporal offset between them. Our goal is to recover the parameters α and Δt in the 1D affine transformation. We address this problem by exploiting a novel projective-invariant descriptor based on the cross ratio to obtain the matched trajectory points between the two input videos.

So far, numerous video synchronization methods have been presented in the previous works, which are mainly classified into two categories: intensity-based ones [9–13] and feature-based ones [14–29]. The intensity-based methods usually rely on colors, intensities, or intensity gradients to achieve the temporal synchronization of overlapping videos. The feature-based methods, however, rely on some detected features like trajectory points.

Among the feature-based video synchronization methods, the trajectory-based ones are one of the most popular categories [19–29]. These methods generally use some epipolar geometry or homography information among different viewpoints for the purpose of exploiting the matched trajectory points or time pairs between (or among) input videos [21, 22, 26, 28]. For example, Padua F, et al., introduced a concept of the timeline and presented a linear sequence-to-sequence alignment algorithm in [21]. They first performed the epipolar geometric constraint (EGC) on each trajectory point and then searched the matched trajectory points (or time pairs). In [22], a rank constraint-based video synchronization method was presented. Owing to the proposed rank constraint, it was not necessary to compute the fundamental matrix between the two views.

Alternatively, some video synchronization methods were proposed by first constructing a projective or multi-view invariant descriptor for each trajectory point and then finding the matched trajectory points or time pairs between the two videos based on the similarities among those descriptors. For example, Liu and Yang [23] proposed a video synchronization method based on the trajectory curvature (TC), where TCs were adopted as the descriptors and a scale-space scoring method was used to match the trajectory points. Since the curvature-based descriptor is not full-affine or projective invariant, this method only works for the videos that are captured from the slightly different viewpoints. In [24] and [25], two projective-invariant representation based video synchronization methods were presented based on the cross ratio of five coplanar points and four collinear points, respectively. These methods were shown to recover both spatial and temporal alignments with good precision and efficiency for the planar scenes.

However, the five coplanar points or the four collinear points were constructed by using the neighbors of each current trajectory point in [24] and [25]. This leads to the following two facts: (1) In the case of the planar or locally planar trajectories, each trajectory point and its neighbors on the trajectory are coplanar, which is not really true for the scenes containing the non-planar motions. As a result, such methods generally work well for the specific scenes containing planar motions rather than the general non-planar motions; (2) During the computation of the cross ratio, only the local spatial information from the trajectories is

¹ In fact, the proposed method does not require any priori information about the temporal relationships between the input sequences. Other temporal misalignment models could also be used in our proposed method. Similar to that in [21], what is required in the proposed method is that the video cameras are stationary, with fixed (but unknown) intrinsic and extrinsic parameters..

employed. The global spatial information of the trajectory point related to the whole scene is easily ignored in these methods. In another words, these methods may achieve some undesirable results or even fail when the trajectories containing numerous segments with the largely similar shapes.

In this paper, a set of five points are first constructed for each trajectory point in each input video, where a trajectory point is used as the reference point and other four fixed points are induced from the background scene. Their corresponding original points in the 3D real-world are *co-planar* even if the motions of the targets in the scene are non-planar. Then a projective-invariant descriptor is presented for each trajectory point by using the cross ratio induced from the five co-planar points. Finally, a novel synchronization method is proposed to synchronize the multiple videos.

Owing to the proposed five-coplanar-points structure, the proposed video synchronization method works well for the scenes containing the more general non-planar motions as well as the planar motions. Moreover, the spatial information of each trajectory point related to the whole background scene is also considered via computing the cross ratio based on the four fixed points from that scene in the proposed synchronization method. This will further improve the synchronization precision. Experimental results demonstrate the validity of the proposed method.

The main contributions in this paper are listed as follows.

- (1) A novel trajectory-based video synchronization method is proposed, where the proposed projective-invariant descriptor and the epipolar geometric constraint are jointly engaged, rather than just the geometric constraint in [21] or the projective invariant representation in [24, 25].
- (2) A robust and distinctive projective-invariant descriptor is presented for each trajectory point, which consists of multiple cross ratio values rather than single cross ratio value. This greatly improves the matching accuracy of the trajectory points.
- (3) A novel five-coplanar-points structure is constructed for each trajectory point during the computation of the cross ratio. In such structure, the proposed descriptor works well for the scenes with the planar or the general non-planar motions.

The rest of this paper is organized as follows: Sect. 2 describes the proposed projective-invariant descriptor in details. In Sect. 3, the proposed video synchronization method is elaborated. The experimental results and the conclusions are given in Sects. 4 and 5, respectively.

2 Projective-Invariant Description

In this section, we will present a novel method to construct a robust and distinctive projective-invariant descriptor for each trajectory point in the input videos. Firstly, we will briefly introduce the computation of the cross ratio of the five coplanar points. Then we will introduce a five-points construction method for each trajectory point such that their original points in the 3D real-world are coplanar even if the trajectory of the moving target in the scene is non-planar.

Mathematically, if the five coplanar points $\{\mathbf{p}_i | i = 1, 2, 3, 4, 5\}$ are expressed in the homogeneous coordinates (i.e., $\mathbf{p}_i = [\lambda x_i, \lambda y_i, \lambda]^T$), their cross ratio can be computed by [24]

$$\Gamma(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4, \mathbf{p}_5) = \frac{|m_{125}| |m_{134}|}{|m_{124}| |m_{135}|} \quad (2)$$

where m_{ijk} is the 3×3 matrix with $\mathbf{p}_i$, $\mathbf{p}_j$ and $\mathbf{p}_k$ as columns, and $|m|$ is the determinant of the matrix m . Here, $\mathbf{p}_1$ is viewed as the reference point of the cross ratio.

Based on the cross ratio, we can construct a projective-invariant descriptor for each trajectory point to establish the frame correspondences between the two input videos. However, *for each trajectory point to be considered, how to construct a set of five points from every input video to compute its cross ratio such that their corresponding original points in the 3D real-world are coplanar* is still an open issue. In addition, there is no ordering ambiguity for the five points under any projective transformation, considering that the cross ratio values depend on the ordering of the five points.

In this paper, the proposed five-points set consists of a moving point from the trajectories and the four fixed points from the background image, where the ordering ambiguity is avoided because of four of the five points are fixed. Moreover, the trajectory point will be taken as the reference point during the computation of the cross ratio. In this case, the remaining problem is how to construct the four fixed points from the background image.

Suppose there are two input videos, denoted by V and V' , respectively. V is the reference video and V' is the video to be synchronized. I and I' are their background images, and $\{(\mathbf{b}_i, \mathbf{b}'_i) | i = 1, 2, \dots, n_0; \mathbf{b}_i \in I, \mathbf{b}'_i \in I'\}$ are a set of matched feature points between the two background images. $\{\mathbf{p}_i | i = 1, 2, \dots, n_1\}$ and $\{\mathbf{p}'_j | j = 1, 2, \dots, n_2\}$ are the trajectory points extracted from the two input videos. $\mathbf{e}$ and $\mathbf{e}'$ are the epipolar points in the two views, and $\mathbf{F}$ denotes the fundamental matrix between the two views. $\mathbf{C}$ and $\mathbf{C}'$ are supposed to be the centers of two cameras. The four fixed points are then constructed as follows.

First, three pairs of the matched feature points $\{(\mathbf{b}_i, \mathbf{b}'_i) | i = 1, 2, 3; \mathbf{b}_i \in I, \mathbf{b}'_i \in I'\}$ are randomly selected from the background images of the two input videos.

Secondly, define a line $\mathbf{l}_1$ passing through points $\{\mathbf{b}_1, \mathbf{b}_2\}$ and a line $\mathbf{l}_2$ passing through points $\{\mathbf{e}, \mathbf{b}_3\}$ in the image I . A point $\mathbf{b}_c$ is obtained as the intersection of lines $\mathbf{l}_1$ and $\mathbf{l}_2$. Similarly, two lines $\mathbf{l}'_1$ and $\mathbf{l}'_2$ are defined through points $\{\mathbf{b}'_1, \mathbf{b}'_2\}$ and points $\{\mathbf{e}', \mathbf{b}'_3\}$ in the image I' , respectively, and a point $\mathbf{b}'_c$ is also obtained as the intersection of lines $\mathbf{l}'_1$ and $\mathbf{l}'_2$.² Obviously, $\mathbf{b}_c$ and $\mathbf{b}'_c$ are also a pair of matched feature points. The four points $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_c, \mathbf{e}\}$ in the video V and their corresponding points $\{\mathbf{b}'_1, \mathbf{b}'_2, \mathbf{b}'_c, \mathbf{e}'\}$ in the video V' are constructed, respectively.

Next, we will prove their original points $\{\mathbf{P}_{ij}, \mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_c, \mathbf{E}\}$ of the five points $\{\mathbf{p}_i, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_c, \mathbf{e}\}$ or $\{\mathbf{p}'_j, \mathbf{b}'_1, \mathbf{b}'_2, \mathbf{b}'_c, \mathbf{e}'\}$ are coplanar in the 3D real-world if the trajectory points $\mathbf{p}_i$ and $\mathbf{p}'_j$ are a pair of the matched ones.

As shown in Fig. 1, the three original points $\{\mathbf{P}_{ij}, \mathbf{B}_1, \mathbf{B}_2\}$ are supposed to be on a plane π_1 . The original point $\mathbf{E}$ corresponding to the epipolar points $(\mathbf{e}, \mathbf{e}')$ on the plane π_1 is determined as the intersection of the line passing through points $\{\mathbf{C}, \mathbf{C}'\}$ on the plane π_1 . That means the four points $\{\mathbf{P}_{ij}, \mathbf{B}_1, \mathbf{B}_2, \mathbf{E}\}$ are all on the plane π_1 and there is no doubt that they are coplanar.

$\mathbf{B}_3$ is not sure to be on the plane π_1 , while $\mathbf{B}_c$ is on the plane π_1 . As discussed previously, $\{\mathbf{b}_1, \mathbf{b}'_1\}$ are a pair of the matched feature points, and $\mathbf{B}_1$ is their original point in the 3D-real world. Also, $\{\mathbf{b}_2, \mathbf{b}'_2\}$ are a pair of the matched feature points, and $\mathbf{B}_2$ is their original point. Correspondingly, $\{\mathbf{l}_1, \mathbf{l}'_1\}$ are a pair of the matched lines and $\mathbf{L}_1$ is their original one, which passes through the points $\mathbf{B}_1$ and $\mathbf{B}_2$ and is obviously on the plane π_1 . The three points $\{\mathbf{C}, \mathbf{b}_1, \mathbf{B}_1\}$ are collinear, and the three points $\{\mathbf{C}, \mathbf{b}_2, \mathbf{B}_2\}$ are also collinear. A plane π_2 , which is not displayed in the Fig. 1, is determined by the two lines. Accordingly, the line $\mathbf{L}_1$

² The three pairs of matched feature points $(\mathbf{b}_i, \mathbf{b}'_i)$, $i = 1, 2, 3$ will be selected again to obtain the intersections $\mathbf{b}_c$ and $\mathbf{b}'_c$ if one of the following cases appear: (1) $\mathbf{l}_1$ is parallel to $\mathbf{l}_2$; (2) $\mathbf{l}'_1$ is parallel to $\mathbf{l}'_2$; (3) $\mathbf{b}_1, \mathbf{b}_2, \mathbf{e}$ are collinear; (4) $\mathbf{b}'_1, \mathbf{b}'_2, \mathbf{e}'$ are collinear.

Fig. 1 Coplanarity of the points $\{P_{ij}, B_1, B_2, B_c, E\}$. Definitions of those symbols in the figure are seen in the text body

is also on the plane π_2 . Similarly, the three points $\{C', b'_1, B'_1\}$ are collinear, and the three points $\{C', b'_2, B'_2\}$ are also collinear. A plane π'_2 is also determined by the two lines, and the line L_1 is also on the plane π'_2 . Therefore, L_1 is the interline of the two planes $\{\pi_2, \pi'_2\}$. b_c and b'_c are a pair of the matched points and are on the lines L_1 and L'_1 , respectively. Therefore, their original point B_c in the 3D real-world will be on the plane π_2 as well as on the plane π'_2 . That is to say, the point B_c is on the interline, i.e., L_1 , of the two planes $\{\pi_2, \pi'_2\}$ and is on the plane π_1 . This indicates that the points $\{P_{ij}, B_1, B_2, B_c, E\}$ are coplanar. It should be also noted that the five points $\{P_{ij}, B_1, B_2, B_c, E\}$ are still coplanar, although the plane π_1 and the point E may vary with the moving point P_{ij} .

Having constructed the above four points $\{b_1, b_2, b_c, e\}$, the cross ratio $\Gamma(p_i, b_1, b_2, b_c, e)$ for each trajectory point $p_i (i = 1, 2, \dots, n_1)$ in the video V can be computed by using Eq. (2). Similarly, with the four points $\{b'_1, b'_2, b'_c, e'\}$, the cross ratio $\Gamma(p'_j, b'_1, b'_2, b'_c, e')$ for each trajectory point $p'_j (j = 1, 2, \dots, n_2)$ in the video V' can also be computed by using Eq. (2). Moreover, $\Gamma(p_i, b_1, b_2, b_c, e)$ and $\Gamma(p'_j, b'_1, b'_2, b'_c, e')$ will be the same ideally, if p_i and p'_j are a pair of the matched trajectory points between the two input videos. Accordingly, after the two sets of fixed four points $\{b_1, b_2, b_c, e\}$ and $\{b'_1, b'_2, b'_c, e'\}$ are individually constructed in the videos V and V' , the matched trajectory points from the two input videos can be simply determined by using their cross ratio values as in [24].

In order to further improve the robustness and accuracy of the subsequent matching of trajectory points, we will use multiple cross ratio values, rather than just single cross ratio value, to describe each trajectory point in this paper. To do so, we will first randomly select three pairs of the matched feature points $(b_i, b'_i), i = 1, 2, 3$ for N times, which is set to 4 empirically. For each time, we obtain a cross ratio value for each trajectory point. Then we construct a projective-invariant descriptor $\mathbf{d} \in R^N$ of dimension N for each trajectory point, each element of which is defined as the cross ratio value computed above. Although its simplicity, the presented descriptor greatly improves the correct matching rate of the proposed method. In summary, the process of building the projective-invariant descriptor is described in Algorithm 1.

Algorithm 1. Construction of projective-invariant descriptors for trajectory points.

Input: Trajectory points $\{\mathbf{p}_i | i = 1, 2, \dots, n_1\}$ and $\{\mathbf{p}'_j | j = 1, 2, \dots, n_2\}$, matched background feature points $\{(\mathbf{b}_i, \mathbf{b}'_i) | i = 1, 2, \dots, n_0; \mathbf{b}_i \in I, \mathbf{b}'_i \in I'\}$, epipolar points $\mathbf{e}$ and $\mathbf{e}'$.

Output: Descriptors $\{\mathbf{d}_i \in R^N | i = 1, 2, \dots, n_1\}$ and $\{\mathbf{d}'_j \in R^N | j = 1, 2, \dots, n_2\}$.

For $n = 1 : N$

1. Randomly select three pairs of matched feature points $(\mathbf{b}_i, \mathbf{b}'_i), i = 1, 2, 3$ from the background images I and I' .
2. In the background image I ,
 - (2.1) compute the line $\mathbf{l}_1$ by using points $\mathbf{b}_1$ and $\mathbf{b}_2$ as $\mathbf{l}_1 = \mathbf{b}_1 \times \mathbf{b}_2$, where $\times$ denotes the cross product of two vectors.
 - (2.2) compute the line $\mathbf{l}_2$ by using points $\mathbf{e}$ and $\mathbf{b}_3$ as $\mathbf{l}_2 = \mathbf{e} \times \mathbf{b}_3$;
 - (2.3) compute the intersection of lines $\mathbf{l}_1$ and $\mathbf{l}_2$ as $\mathbf{b}_i = \mathbf{l}_1 \times \mathbf{l}_2$.
 - (2.4) For each trajectory point $\mathbf{p}_i (i = 1, 2, \dots, n_1)$, compute the cross ratio $\Gamma(\mathbf{p}_i, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_i, \mathbf{e})$ by using Eq. (2) and set the n -th element $\mathbf{d}_i(n)$ of the descriptor $\mathbf{d}_i$ as $\mathbf{d}_i(n) = \Gamma(\mathbf{p}_i, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_i, \mathbf{e})$, where $\mathbf{p}_i$ is taken as the reference point.
3. Similarly, in the background image I' ,
 - (3.1) compute the line $\mathbf{l}'_1$ by $\mathbf{l}'_1 = \mathbf{b}'_1 \times \mathbf{b}'_2$;
 - (3.2) compute the line $\mathbf{l}'_2$ by $\mathbf{l}'_2 = \mathbf{e}' \times \mathbf{b}'_3$.
 - (3.3) compute the intersection point $\mathbf{b}'_i$ by $\mathbf{b}'_i = \mathbf{l}'_1 \times \mathbf{l}'_2$;
 - (3.4) For each trajectory point $\mathbf{p}'_j (j = 1, 2, \dots, n_2)$, compute the cross ratio $\Gamma(\mathbf{p}'_j, \mathbf{b}'_1, \mathbf{b}'_2, \mathbf{b}'_i, \mathbf{e}')$ by using Eq. (2) and set the n -th element $\mathbf{d}'_j(n)$ of the descriptor $\mathbf{d}'_j$ as $\mathbf{d}'_j(n) = \Gamma(\mathbf{p}'_j, \mathbf{b}'_1, \mathbf{b}'_2, \mathbf{b}'_i, \mathbf{e}')$, where $\mathbf{p}'_j$ is taken as the reference point,

End for

From the construction of the five-points set, it is worth noting that their original points in the 3D real-world of those five points remain coplanar even if the motions of the moving targets in the scene are planar or non-planar. As a result of that, the proposed projective-invariant descriptor still works well for the scenes with non-planar motions. Moreover, owing to the four fixed points from the background scene, the spatial position information of each trajectory point related to the whole background scene is integrated into the proposed descriptor. This will further improve the accuracy in the subsequent matching of the trajectory points.

3 Proposed Video Synchronization Method

As shown in Fig. 2, the proposed method contains three parts: (I) Trajectory extraction and background image matching; (II) Trajectory point description and matching; (III) Estimation of temporal relationship.

For step I, the backgrounds of input videos are simply assumed to be static and the frame-difference-based methods [30, 31] are performed on the reference video V and the video V' to be synchronized, respectively. Some other methods could also be employed here, which are beyond the scope of this paper. Then two background images I and I' , and two sets of trajectory points³ $\{\mathbf{p}_i | i = 1, 2, \dots, n_1\}$ and $\{\mathbf{p}'_j | i = 1, 2, \dots, n_2\}$ are, respectively, extracted from the two input videos. Here, i denotes the frame index of the video V , and j denotes the frame index of the video V' . n_1 and n_2 are the total numbers of the frames contained in the two videos, respectively. Consequently, we can obtain the matched feature points $\{(\mathbf{b}_i, \mathbf{b}'_i) | i = 1, 2, \dots, n_0; \mathbf{b}_i \in I, \mathbf{b}'_i \in I'\}$ by using some image matching methods such as SIFT [32] or ASIFT [33].

The synchronization between the two videos V and V' , or the estimation of the parameters in the temporal transformation model in Eq. (1), can be achieved by finding the matched trajectory points $(\mathbf{p}_i, \mathbf{p}'_j)$ or frame pairs (i, j) between the two videos.

³ Here, we assume that there is only one moving object in the scene for simplicity. The proposed method can also be applied to the case of multiple moving objects in the scene. See Subsection 4.2.

Fig. 2 Diagram of the proposed method

In the following subsections, we will describe step II (i.e., trajectory point description and matching) and step III (i.e., temporal relationship estimation) of our proposed method in details.

3.1 Trajectory Point Description and Matching

The procedure of trajectory point description and matching is described as follows:

- (1) Given the matched feature points $\{(\mathbf{b}_i, \mathbf{b}'_i) | i = 1, 2, \dots, n_0; \mathbf{b}_i \in I, \mathbf{b}'_i \in I'\}$, compute the epipolar geometry between the two views by using the simple-normalized eight-point algorithm [34] or some robust epipolar geometry estimation methods [35, 36] and obtain the Fundamental matrix $\mathbf{F}$ as well as the epipolar points $(\mathbf{e}, \mathbf{e}')$ in the two views.
- (2) Compute the projective-invariant descriptor $\mathbf{d}_i$ ($i = 1, 2, \dots, n_1$) for each trajectory point $\mathbf{p}_i$ ($i = 1, 2, \dots, n_1$) in the video V and the projective-invariant descriptor $\mathbf{d}'_j$ ($j = 1, 2, \dots, n_2$) for each trajectory point $\mathbf{p}'_j$ ($j = 1, 2, \dots, n_2$) in the video V' based on Algorithm 1, respectively.
- (3) Obtain the initially matched trajectory point pairs set $M_T^{(1)} = \{(\mathbf{p}_i, \mathbf{p}'_j) | \mathbf{p}_i \in V, \mathbf{p}'_j \in V'\}$ via using the nearest neighbor distance ratio based method [37]. More specifically, the trajectory points $\mathbf{p}_i$ and $\mathbf{p}'_j$ are treated as a pair of the matched ones if they satisfy

$$\frac{D(\mathbf{d}_i, \mathbf{d}'_j)}{D(\mathbf{d}_i, \mathbf{d}'_{j'})} \leq th1 \quad (3)$$

where $D(\cdot)$ denotes the Euclidean distance between two descriptor vectors. The points $\mathbf{p}'_j$ and $\mathbf{p}'_{j'}$ are, respectively, the first and second nearest ones to the point $\mathbf{p}_i$ according to the Euclidean distances of their descriptors. $th1$ is a threshold with a value of 0.85 in this paper.

- (4) Employ the epipolar geometric constraint to remove some obvious mismatches in the set $M_T^{(1)}$ and obtain a refined set $M_T^{(2)}$. For each pair of the matched points $(\mathbf{p}_i, \mathbf{p}'_j)$ in the set $M_T^{(1)}$, determine the epipolar line $\mathbf{l}'_{\mathbf{p}_i}$ of the point $\mathbf{p}_i$ in the video V' and then compute the distance $d(\mathbf{p}'_j, \mathbf{l}'_{\mathbf{p}_i})$ of the point $\mathbf{p}'_j$ to the line $\mathbf{l}'_{\mathbf{p}_i}$. Similarly, determine the epipolar line $\mathbf{l}_{\mathbf{p}'_j}$ of the point $\mathbf{p}'_j$ in the video V and compute the distance $d(\mathbf{p}_i, \mathbf{l}_{\mathbf{p}'_j})$ of the

point $\mathbf{p}_i$ to the line $\mathbf{l}_{\mathbf{p}'_j}$. If $d(\mathbf{p}'_j, \mathbf{l}_{\mathbf{p}_i})$ and $d(\mathbf{p}_i, \mathbf{l}_{\mathbf{p}'_j})$ satisfy Eq. (4), the points $(\mathbf{p}_i, \mathbf{p}'_j)$ are viewed as a pair of the correctly matched ones. Otherwise, the points $(\mathbf{p}_i, \mathbf{p}'_j)$ are regarded as a pair of the mismatched ones and are removed from $M_T^{(1)}$. Then a refined set $M_T^{(2)}$ is obtained.

$$d(\mathbf{p}'_j, \mathbf{l}'_i) \leq th2 \& d(\mathbf{p}_i, \mathbf{l}_j) \leq th2 \quad (4)$$

here the threshold $th2$ is adaptively set as

$$th2 = \frac{1}{2n_0} \sum_{i=1}^{n_0} (d(\mathbf{b}_i, \mathbf{l}_{\mathbf{b}'_i}) + d(\mathbf{b}'_i, \mathbf{l}_{\mathbf{b}_i})) \quad (5)$$

where $\mathbf{l}_{\mathbf{b}_i}$ denotes the epipolar line of the point $\mathbf{b}_i$ in the video V' , and $\mathbf{l}_{\mathbf{b}'_i}$ denotes the epipolar line of the point $\mathbf{b}'_i$ in the video V . n_0 is the total number of the matched feature points between the two background images.

- (5) Extract the frame indices from the matched points $(\mathbf{p}_i, \mathbf{p}'_j)$ in the set $M_T^{(2)}$ and obtain the matched frame pairs set $M_F^{(1)} = \{(i, j) | (\mathbf{p}_i, \mathbf{p}'_j) \in M_T^{(2)}\}$.

3.2 Estimation of Temporal Relationship

We can directly perform the Random Sample Consensus (RANSAC) [38] algorithm on the matched frame pairs set $M_F^{(1)}$ obtained in the previous subsection to estimate the temporal transformation model between the two input videos. However, there are still many mismatches in the set $M_F^{(1)}$ because of the computational errors. This will reduce the final estimation precision, although RANSAC is robust to the mismatches or “outliers” to some extent.

As discussed in the earlier Sect. 2, three pairs of points $(\mathbf{b}_i, \mathbf{b}'_i)$, $i = 1, 2, 3$ are randomly selected from the matched background feature points when the cross ratio for each trajectory point is computed. Accordingly, different sets of matched frame pairs $M_F^{(1)}$ are obtained when the trajectory point description and matching method introduced in the previous subsection is performed on the same input videos by several times. Despite of that, we also find that the correctly matched frame pairs generally appear in the set $M_F^{(1)}$ with higher probability than those mismatched ones.

Based on such observation, we present a simple but effective temporal parameter or model estimation method in this paper, which is described as follows.

- (1) Perform the trajectory point description and matching method introduced in the previous subsection on the input videos several times (e.g., S times) and obtain S sets of the matched frame pairs $\{M_F^{(s)} | s = 1, \dots, S\}$ for the same pair of input videos.
- (2) Define a score matrix M_C of orders $n_1 \times n_2$ to count the total times $T_{i,j}$ of each matched frame pair (i, j) appearing in those sets $\{M_F^{(s)} | s = 1, \dots, S\}$ with $M_C(i, j) = T_{i,j}$. $M_C(i, j)$ denotes the (i, j) -th entry of the matrix M_C . A higher value of $M_C(i, j)$ means that the i -th frame in the video V and the j -th frame in the video V' are more possible to be a pair of the correctly matched frames.
- (3) Compute the final matched frame pairs set M_F as

$$M_F = \{(i, j) | M_C(i, j) \geq \max(M_C(:)) \times r\} \quad (6)$$

where $\max(M_C(:))$ denotes the maximum value in matrix M_C . r is experimentally set to 0.4 in this paper.

- (4) Estimate the parameters α and Δt of the temporal transformation model in Eq. (1) by performing RANSAC on the matched frame pairs in M_F .

In summary, the pseudo code of the proposed video synchronization method is presented in Algorithm 2.

Algorithm 2. Pseudo code of the proposed video synchronization method.

Input: Reference video V and the video V' to be synchronized

Output: Temporal parameters α and Δt

Begin

1. Extract the trajectory points $\{\mathbf{p}_i | i=1,2,\dots,n_1\}$, $\{\mathbf{p}'_j | j=1,2,\dots,n_2\}$ and background images I and I' of the two input videos.
2. Compute the matched feature points $\{(\mathbf{b}_i, \mathbf{b}'_i) | i=1,2,\dots,n_0; \mathbf{b}_i \in I, \mathbf{b}'_i \in I'\}$, and then compute the fundamental matrix $\mathbf{F}$ and the epipolar points $(\mathbf{e}, \mathbf{e}')$ in the two views.
3. Compute the matched trajectory points or frame pairs:

For $s=1:S$

- 3.1. Construct the projective-invariant descriptors $\{\mathbf{d}_i \in R^N | i=1,2,\dots,n_1\}$ and $\{\mathbf{d}'_j \in R^N | j=1,2,\dots,n_2\}$ for the trajectory points $\{\mathbf{p}_i | i=1,2,\dots,n_1\}$ and $\{\mathbf{p}'_j | j=1,2,\dots,n_2\}$ by using Algorithm 1, respectively.
- 3.2. Compute the initial matched trajectory point pairs set $M_T^{(1)}$.
- 3.3. Remove some mismatched trajectory point pairs in the set $M_T^{(1)}$ by using the epipolar geometric constraint and obtain the refined set $M_T^{(2)}$.
- 3.4. Construct the matched frame pairs set $M_F^{(s)}$.
- 3.5. Update the score matrix M_C with

$$M_C(i, j) = \begin{cases} M_C(i, j) + 1, & \text{if } (i, j) \in M_F^{(s)} \\ M_C(i, j), & \text{otherwise} \end{cases} \quad (7)$$

End for

4. Obtain the final matched frame pairs set M_F by using Eq. (6).
5. Estimate the parameters α and Δt in the temporal transformation model by performing RANSAC on the matched frame pairs in the set M_F .

End

4 Experiments and Analysis

In this section, two sets of experiments are conducted to demonstrate the validity of the proposed video synchronization method. First, the impacts of some parameters are discussed by using a pair of videos. Secondly, several pairs of videos are employed to thoroughly test the effectiveness of the proposed method.

The average temporal alignment error ε_t and the correct matching rate C_r are employed to objectively evaluate different algorithms. Here, ε_t is defined by

$$\varepsilon_t = \frac{1}{n_1} \sum_{i=1}^{n_1} |(\alpha_0 i + \Delta t_0) - (\alpha^* i + \Delta t^*)| \quad (8)$$

where i and n_1 is the frame index and the total number of the frames in the reference video V , respectively. α^* , Δt^* are the estimated parameters for the temporal transformation model between the two input videos. α_0 , Δt_0 are the ground-truth, which are predefined or manually determined in advance. A matched frame pair (i, j) is deemed as a pair of the correct ones if they satisfy

$$\frac{|\alpha_0 i + \Delta t_0 - j|}{\sqrt{\alpha_0^2 + 1}} \leq th3 \quad (9)$$

Table 1 Descriptions of the test input videos

Videos	Number of moving targets	Motion type	Viewpoint angles	Ground truth
<i>Car</i> ^a	1	Planar	Small	$\alpha_0 = 1, \Delta t_0 = 5$
<i>Forkroad</i> ^b	Multiple	Planar	Large	$\alpha_0 = 1, \Delta t_0 = 50$
<i>Footbridge</i>	Multiple	Planar	Large	$\alpha_0 = 2, \Delta t_0 = 8$
<i>Toyplay</i>	1	Non-planar	Large	$\alpha_0 = 1, \Delta t_0 = 0$

^aThese videos are downloaded from <http://www.wisdom.weizmann.ac.il/~vision/VideoAnalysis/Demos/Seq2Seq/Seq2Seq.html>

^bThese videos are downloaded from <http://www.cvg.reading.ac.uk/PETS2001/pets2001-dataset.html>

Fig. 3 Background images and trajectory points extracted from the test input videos. **a** *Car*; **b** *Forkroad*; **c** *Footbridge*; **d** *Toyplay*. The images in the top row are the background images of the reference videos, and the images in the bottom row are the background images of the videos to be synchronized. The red points are the trajectory points extracted from the test videos

where $th3$ is a threshold and is experimentally set to 0.5 in all of the experiments. C_r is defined as the ratio between the number of the correct matches and the total number of the matches in the set M_F . Theoretically, smaller values of ε_t and higher values of C_r indicate the better performance of a method.

Table 1 lists several pairs of videos employed in this paper, and Fig. 3 illustrates their background images and the trajectory points extracted from these videos. It should be noted that the motions of the moving targets in the “*Toyplay*” videos are non-planar.

4.1 Impacts of Some Parameters

In this subsection, we will employ Videos *Car* in Table 1 to test the impacts on the system performance of some parameters in the proposed method, including the dimension of the descriptor N in Algorithm 1 and the parameter S in Algorithm 2.

The ε_t and C_r curves with the two parameters are provided in Fig. 4. Figure 4a, b indicate that the performance of the proposed method varies continuously with N and achieves the best when N is set to 4. As shown in Fig. 4c, d, the values of ε_t decrease dramatically and the values of C_r increase obviously with S at the beginning. However, the performance of the proposed method remains nearly unchanged when S is around 100. To facilitate the following experiments, we will set N to 4 and S to 100, respectively.

Fig. 4 Performance of the proposed method with parameters N and S , respectively. **a** ε_t curve with N ; **b** C_r curve with N ; **c** ε_t curve with S ; **d** C_r curve with S

Fig. 5 Temporal alignment results on Videos *Car*. **a–g** are the alignment results obtained by RC, EGC, SPIR, MT, TC, MB and the proposed method, respectively. **a** $RC\ j = 0.9935 * i + 5.3589$, **b** $EGC\ j = 0.9979 * i + 5.9948$, **c** $SPIR\ j = 1.0034 * i + 6.7194$, **d** $MT\ j = 1.015 * i + 1.453$, **e** $TC\ j = 1.011 * i - 0.4463$, **f** $MB\ j = 1.0118 * i + 3.9504$, **g** $Proposed\ j = 0.9982 * i + 5.8486$

The experimental results in Fig. 4a, b demonstrate the superiority of multiple cross ratio values (i.e., $N = 4$) employed in the proposed descriptor over the single cross ratio value (i.e., $N = 1$). While the experimental results in Fig. 4c, d also indicate that the performance of the proposed method can be greatly improved by using the score matrix despite its simplicity.

4.2 Validity of the Proposed Method

The four pairs of the input videos in Table 1 are employed to thoroughly demonstrate the validity of the proposed method in this subsection. In addition, some previous methods,

Fig. 6 Temporal alignment results on Videos *Forkroad*. **a–g** are the alignment results obtained by RC, EGC, SPIR, MT, TC, MB and the proposed method, respectively. **a** RC $j = 0.9901 * i + 49.6535$, **b** EGC $j = 0.9932 * i + 51.0856$, **c** SPIR $j = 1 * i + 48$, **d** MT $j = 1.3045 * i + 17.6967$, **e** TC $j = 0.8898 * i + 47.5864$, **f** MB $j = 1.0124 * i - 15.1112$, **g** Proposed $j = 0.9945 * i + 50.825$

Fig. 7 Temporal alignment results on Videos *Footbridge*. **a–g** are the alignment results obtained by RC, EGC, SPIR, MT, TC, MB and the proposed method, respectively. **a** RC $j = 1.999 * i + 9.0093$, **b** EGC $j = 1.9971 * i + 8.2176$, **c** SPIR $j = 2.042 * i + 4.0875$, **d** MT $j = 1.9989 * i + 6.1125$, **e** TC $j = 1.9532 * i + 15.4884$, **f** MB $j = 2.1459 * i + 7.9369$, **g** Proposed $j = 2 * i + 8$

including RC [22], EGC [21], SPIR [24], MT [25], TC [23], and MB [26], are performed on these videos for comparisons.

The estimation results obtained by different methods on the four pairs of the input videos are shown in Figs. 5, 6, 7 and 8, respectively. These figures demonstrate that the estimated temporal parameters α_* and Δt_* obtained by the proposed method are closer to their corresponding ground truth values than those computed by other methods. Particularly, the estimated temporal parameters α_* and Δt_* obtained by the proposed method are the same as their ground truth values for Videos *Footbridge*.

The values of metrics ε_t and C_r obtained by different methods are provided in Table 2 and Fig. 9, respectively, which also verify the obvious superiority of our proposed method over others. In addition, the total computing time of different methods is also provided in Table 3, which indicates that the computational efficiency of the proposed method is

Fig. 8 Temporal alignment results on Videos *Toyplay*. **a–g** are the alignment results obtained by RC, EGC, SPIR, MT, TC, MB and the proposed method, respectively. **a** $RC\ j = 0.9763 * i + 1.9042$, **b** $EGC\ j = 0.9709 * i + 1.8572$, **c** $SPIR\ j = 0.9229 * i - 1.9093$, **d** $MT\ j = 0.5386 * i + 13.4569$, **e** $TC\ j = 0.9301 * i + 0.9917$, **f** $MB\ j = 0.822 * i + 8.3996$, **g** $Proposed\ j = 0.9998 * i + 0.1022$

Table 2 Average temporal alignment errors ε_t obtained by different methods

Input videos	RC	EGC	SPIR	MT	TC	MB	Proposed
<i>Car</i>	1.2998	0.4673	2.5755	1.8778	2.7085	2.0827	0.3992
<i>Forkroad</i>	2.9956	1.0507	2	3	3.6848	61.7827	0.4632
<i>Footbridge</i>	0.9076	0.15227	15.4043	75.2358	29.4139	14.232	0
<i>Toyplay</i>	0.78906	0.0502	6.5737	17.6188	3.345	5.6035	0.0921

Fig. 9 Correct matching rate C_r obtained by different methods. **a** *Car*; **b** *Forkroad*; **c** *Footbridge*; **d** *Toyplay*

Table 3 Total computing time of different methods (in s)

Input videos	RC	EGC	SPIR	MT	TC	MB	Proposed
<i>Car</i>	1.2998	0.4673	2.5755	1.8778	2.7085	2.0827	0.3992
<i>Forkroad</i>	2.9956	1.0507	2	14.9827	8.7494	146.698	0.8809
<i>Footbridge</i>	65.929	3.426	9.621	9.957	391.019	600.698	45.739
<i>Toyplay</i>	4.6	1.033	0.996	1.007	9.52	38.22	9.8554

acceptable. Specifically, for Videos *Car* and *Forkroad*, the proposed method achieves the highest computational efficiency.

Fig. 10 Visual temporal alignment results obtained by RC, EGC, SPIR, MT, TC, MB and the proposed method on Videos *Footbridge*. The first row shows the 31st, 60th, 185th, 195th frames in the reference video. The second to eighth rows show the matched frames computed by RC, EGC, SPIR, MT, TC, MB and the proposed method, respectively

For better comparisons, we illustrate some visual temporal alignment results of these methods on Videos *Footbridge* and *Toyplay* in Figs. 10 and 11, respectively. The first row of Fig. 10 illustrates four representative frames in the reference video of *Footbridge*. The four matched frames in the video to be synchronized computed by RC, EGC, SPIR, MT, TC, MB

Fig. 11 Visual temporal alignment results obtained by RC, EGC, SPIR, MT, TC, MB and the proposed method on Videos *Toyplay*. The first row shows the 8th, 33rd, 71st, 87th frames in the reference video. The second to eighth rows show the matched frames in the video to be synchronized computed by RC, EGC, SPIR, MT, TC, MB and the proposed method, respectively. In each of these sub-figures, the circle denotes the moving target (i.e., playing toy) in the scene, and the red cross (if existed) denotes the centroid location of the playing toy in the corresponding frame of the reference video

and the proposed method are shown in the rest rows. A zoomed version of the representative image block in each frame is also provided in a small window. By comparing these zoomed

blocks, there is no doubt that the proposed method obtains the highest temporal alignment precisions among those baselines above.

For example, as shown in the small window of the 60th frame (row 1, column 2 in Fig. 10) in the reference video, the right foot is behind the left foot. While in the matched frame computed by SPIR (row 4, column 2 in Fig. 10), the right foot is ahead the left foot. In the 185th frame (row 1, column 3 in Fig. 10) of the reference video, the right foot is at the cross of the ground. But in the matched frame computed by MT (row 5, column 3 in Fig. 10), no foot is at the cross of the ground. In contrast, for all of the four frames in the reference video, the proposed method can obtain the correctly matched ones, as shown in the last row of Fig. 10. Similar conclusions can also be made from Fig. 11. Those indicate that our proposed method has achieved better performance for the scenes with both the 3D non-planar and planar motions. This mainly gives credits to the proposed projective invariant descriptor, especially the proposed five-points structure during the computation of the cross ratio.

5 Conclusion

In this paper, we first propose a novel method to construct a robust and distinctive projective-invariant descriptor for each trajectory point by using the cross ratio of the proposed five coplanar points, which consist of a moving point from the trajectories and the four fixed points from the background images. More specifically, the five points remain coplanar in the 3D real-world even if the target motions are non-planar. Then we present a novel video synchronization framework that integrates the proposed projective-invariant descriptor and the constructed five-points structure jointly. The proposed method performs well for the scenes containing the general non-planar motions as well as the planar motions on account of the above novelties. Experimental results demonstrate that the proposed method significantly outperforms some state-of-the-art baselines for those videos being captured under different conditions, including different frame rates, wide baselines, multiple moving objects, planar or non-planar trajectories. In the future work, we will apply the presented approach in this paper to the automated multi-camera surveillance and the co-saliency detection [39–41].

Acknowledgements This work is supported by the National Natural Science Foundation of China under Grant No. 61773301, by Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2016JM6008), by the Fundamental Research Funds for the Central Universities under Grant No. JBZ170401.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Caspi Y, Simakov D, Irani M (2006) Feature-based sequence-to-sequence matching. *Int J Comput Vision* 68(1):53–64
2. Wang X (2013) Intelligent multi-camera video surveillance: a review. *Pattern Recogn Lett* 34(1):3–19
3. Wang J, Gao L, Lee YM et al (2016) Target identification of natural and traditional medicines with quantitative chemical proteomics approaches. *Pharmacol Ther* 162:10–22
4. Ali S, Daul C, Galbrun E, Galbrun E, Guillemin F, Blondel W (2016) Anisotropic motion estimation on edge preserving Riesz wavelets for robust video mosaicing. *Pattern Recogn* 51:425–442
5. Zhang D, Han J, Jiang L, Ye S, Chang X (2017) Revealing event saliency in unconstrained video collection. *IEEE Trans Image Process* 26(4):1746–1758

6. Vemulapalli R, Arrate F, Chellappa R (2014) Human action recognition by representing 3D skeletons as points in a lie group. In: IEEE conference on computer vision and pattern recognition, pp 588–595
7. Han J, Chen H, Liu N, Yan C, Li X (2017) CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. *IEEE Trans Cybern*. <https://doi.org/10.1109/TCYB.2017.2761775> (In press)
8. Zhang Q, Wang Y, Levine MD, Yuan X, Wang L (2015) Multisensor video fusion based on higher order singular value decomposition. *Inf Fus* 24:54–71
9. Diego F, Serrat J, López AM (2013) Joint spatio-temporal alignment of sequences. *IEEE Trans Multimed* 15(6):1377–1387
10. Caspi Y, Irani M (2002) Spatio-temporal alignment of sequences. *IEEE Trans Pattern Anal Mach Intell* 24(11):1409–1424
11. Schweiger F, Schroth G, Eichhorn M, Al-Nuaimi A, Cizmeci B, Fahrmar M, Steinbach E (2013) Fully automatic and frame-accurate video synchronization using bitrate sequences. *IEEE Trans Multimed* 15(1):1–14
12. Dai C, Zheng Y, Li X (2006) Accurate video alignment using phase correlation. *IEEE Signal Process Lett* 13(12):737–740
13. Diego F, Ponsa D, Serrat J, Lopez AM (2011) Video alignment for change detection. *IEEE Trans Image Process* 20(7):1858–1869
14. Evangelidis GD, Bauckhage C (2013) Efficient subframe video alignment using short descriptors. *IEEE Trans Pattern Anal Mach Intell* 35(10):2371–2386
15. Zhou F, De la Torre F (2016) Generalized canonical time warping. *IEEE Trans Pattern Anal Mach Intell* 38(2):279–294
16. Pundik D, Moses Y (2010) Video synchronization using temporal signals from epipolar lines. In: European conference on computer vision, pp 15–28
17. Zini L, Cavallaro A, Odono F (2013) Action-based multi-camera synchronization. *IEEE J Emerg Sel Top Circuits Syst* 3(2):165–174
18. Shrestha P, Barbieri M, Weda H, Sekulovski D (2010) Synchronization of multiple camera videos using audio-visual features. *IEEE Trans Multimed* 12(1):79–92
19. Brito DN, Pádua FLC, Pereira GAS, Carceroni RL (2011) Temporal synchronization of non-overlapping videos using known object motion. *Pattern Recogn Lett* 32(1):38–46
20. Pribanic T, Lelas M, Krois I (2015) Sequence-to-sequence alignment using a pendulum. *IET Comput Vision* 9(4):570–575
21. Padua F, Carceroni R, Santos G, Kutulakos K (2010) Linear sequence-to-sequence alignment. *IEEE Trans Pattern Anal Mach Intell* 32(2):304–320
22. Rao C, Gritai A, Shah M, Syeda-Mahmood T (2003) View-invariant alignment and matching of video sequences. In: International conference on computer vision, pp 939–945
23. Liu Y, Yang M, You Z (2012) Video synchronization based on events alignment. *Pattern Recogn Lett* 33(10):1338–1348
24. Nunziati W, Sclaroff S, Bimbo A (2010) Matching trajectories between video sequences by exploiting a sparse projective invariant representation. *IEEE Trans Pattern Anal Mach Intell* 32(3):517–529
25. Wu Y, He X, Nguyen T Q (2013). Subframe video synchronization by matching trajectories. In: International conference on acoustics, speech and signal processing (ICASSP), pp 2277–2281
26. Lu C, Mandal M (2013) A robust technique for motion-based video sequences temporal alignment. *IEEE Trans Multimedia* 15(1):70–82
27. Nunziati W, Sclaroff S, Bimbo AD (2015) An invariant representation for matching trajectories across uncalibrated video streams. In: International conference on image and video retrieval, pp 318–327
28. Lu C, Singh M, Cheng I, Basu A, Mandal M (2011) Efficient video sequences alignment using unbiased bidirectional dynamic time warping. *J Vis Commun Image Represent* 22(7):606–614
29. Cao X, Wu L, Xiao J, Foroosh H, Zhu J, Li X (2010) Video synchronization and its application to object transfer. *Image Vis Comput* 28(1):92–100
30. Singla N (2014) Motion detection based on frame difference method. *Int J Inf Comput Technol* 4(15):1559–1565
31. Comaniciu D, Ramesh V, Meer P (2000) Real-time tracking of non-rigid objects using mean shift. In: Computer vision and pattern recognition, pp 142–149
32. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 60(2):91–110
33. Morel J, Yu G (2009) ASIFT: a new framework for fully affine invariant image comparison. *SIAM J Imag Sci* 2(2):438–469
34. Hartley R, Zisserman A (2003) Multiple view geometry in computer vision. Cambridge University Press, Cambridge
35. Chai J, Ma S (1998) Robust epipolar geometry estimation using genetic algorithm. *Pattern Recogn Lett* 19(9):829–838

36. Chum O, Werner T, Matas J (2004) Epipolar geometry estimation via RANSAC benefits from the oriented epipolar constraint. In: International conference on pattern recognition, pp 112–115
37. Cao Y, Zhang H, Gao Y, Xu X, Guo J (2010) Matching image with multiple local features. In: International conference on pattern recognition, pp 519–522
38. Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM* 24(6):381–395
39. Han J, Cheng G, Li Z, Li Z, Zhang D (2017) A unified metric learning-based framework for co-saliency detection. *IEEE Trans Circuits Syst Video Technol*. <https://doi.org/10.1109/tcsvt.2017.2706264> (In Press)
40. Yao X, Han J, Zhang D, Nie F (2017) revisiting co-saliency detection: a novel approach based on two-stage multi-view spectral rotation co-clustering. *IEEE Trans Image Process* 26(7):3196–3209
41. Zhang D, Meng D, Han J (2016) Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE Trans Pattern Anal Mach Intell* 39(5):865–878

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.