Robust Registration of Long Sport Video Sequence *

GuoJun Liu, XiangLong Tang, Da Sun, and JianHua Huang

Department of Computer Science, Harbin Institute of Technology, China
hitliu@hit.edu.cn

Abstract. Automatic registration plays an important role for a sport analysis system, the automation and accuracy of the registration for a long video sequence can still be an open problem for many practical applications. We propose a novel method to cope with it: (1) Reference frames can be introduced as a transaction of computing homography to map each frame of the imagery to the globally consistent model of the rink, that can reduce the accumulative error of successive registration and make the system more automatic. (2) An more distinctive invariant point feature (SIFT) can be used to provide reliable and robust matching across large range of affine distortion and change in illumination, that can improve the computational precision of homography. Experimental results show that the proposed algorithm is very efficient and effective on video recorded live by the authors in the World Short Track Speed Skating Championships.

1 Introduction

Video-based analysis of sport events is an important tool in analysis of individual players and sport teams, but usually requires many hours of manual work. Recent advances in computer vision can provide help in automating those tasks, such as tracking, annotation, indexing and automatic generation of semantic descriptions. Automatic registration plays an important role for a sport analysis system, it can compensate for camera motions by calculating a planar projective transformation named homography [1]. The automation and accuracy of the registration for a long video sequence can still be an open problem for many practical applications.

The means of video recording can be classified into four types: (1)*Sensors*: many commercial tools [2, 3] are applied successfully to capture and analyze the motion of player in sports using commercially available high-speed high-accuracy measurement system. Due to their limita1tions, these devices are not suitable for studying large-scale motion during a match. In the TRAKUS system [4], the real time acquisition of the player's position is based on the array of microwave receivers which can analyze the signal emitted from special transmitters. These

* Supported by National OlympicScience Foundation (No. 03035) and National Science Foundation (No. 60672090) of China

requirements are hard to fulfill in many regular sports. (2)*Multiple cameras*: Junior et al. [5] develop a real-time distributed system using five static cameras to get the whole scene. however, the placement of cameras and its consistency can limit in other sports. (3)*static cameras*: Pers et al. [6] use two stationary cameras mounted directly above the count and propose a new approach by modeling the radial image distortion more accurately. But the cameras must be placed above the playing count, which is a rigorous condition during regular league or championship matches. (4)*Moving cameras*: Okuma et al. [7] develop a hockey annotation system to automatically analyze hockey scenes that transforms the original sequence in broadcast video to the globally consistent map of the hockey rink using KLT tracker, similarly in [8].

In this paper, we focus on automatic registration as a subsystem of our novel computer vision system for tracking high-speed non-rigid skaters over a large playing area in short track speeding skating competitions. Several important features distinguish the proposed approach from others:

1. Introducing the reference frames as a transition through which each frame can be mapped to the field model in order to reduce the error accumulation of the projection, it's very important for a long video sequence and helpful for improving the precision of the system.
2. An more distinctive invariant point feature (SIFT) can be used to provide reliable and robust matching across large range of affine distortion and change in illumination, that can improve the computational precision of homography.
3. Analyzing the precision of registration ignored by most authors.

The paper is organized as follows: the next section overview the system architecture of our application. Section 3 explores how to automatically compute the mappings that transform each frame to the model of the real rink. Experiments and results are given in section 4 and the conclusion in section 5.

2 Overview of our system

Our special system developed based on computer vision aims to automatically track the movements of sports players (especially skaters) on the large-scale complex and dynamic rink. It will be applied to not only daily training but also the competition. Therefore, a single panning camera is more suitable, it can be mounted at the top auditorium of the stadium as possible as close to the center in order to reduce the projection error. Due to little texture information on the rink unlike [9, 7], zooming will be abandoned in practical application because it can make recording the high-speed target more difficult and enlarge the error of lens distortion. Though the camera center moves by a negligible amount due to the small offset from the camera's optic center, the approximation of pure rotation is indeed sufficient such as proven in [10].

The architecture of our system is shown in Fig. 1, automatic registration as a kernel subsystem affects directly not only the accuracy of the system's

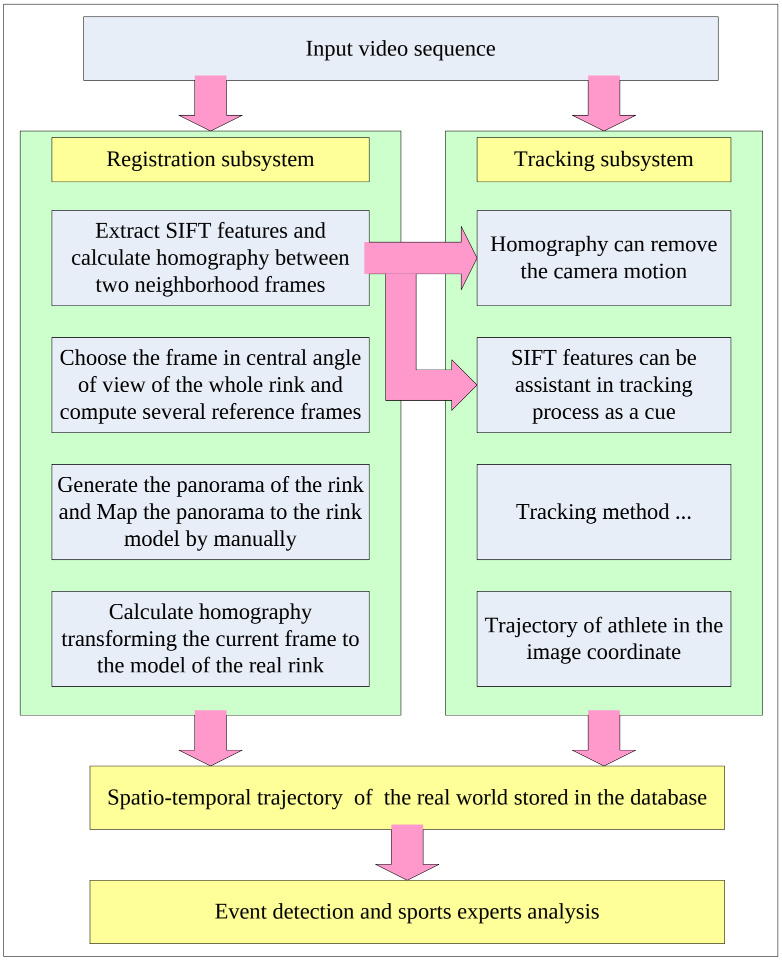

Input video sequence
Registration subsystem
Extract SIFT features and calculate homography between two neighborhood frames
Choose the frame in central angle of view of the whole rink and compute several reference frames
Generate the panorama of the rink and Map the panorama to the rink model by manually
Calculate homography transforming the current frame to the model of the real rink
Tracking subsystem
Homography can remove the camera motion
SIFT features can be assistant in tracking process as a cue
Tracking method ...
Trajectory of athlete in the image coordinate
Spatio-temporal trajectory of the real world stored in the database
Event detection and sports experts analysis

Fig. 1. The architecture of our system.

output but also the tracking performance, it can provide three kinds of outputs to other modules:(1) The homography between two neighborhood frames can remove the camera motion to improve the precision of the motion prediction in the process of tracking. (2) SIFT features extracted can be assistant in tracking process as a cue of target recognition. (3) Another homography transforming the current frame to the model of the real rink is combined with the output of tracking subsystem (the trajectory of athlete in the image coordinate) to obtain the player's spatio-temporal trajectory of the real world, and that can be stored in the database.

3 Automatic registration

The goal of automatic registration for sport applications is transforming positions in the video frame to real-world coordinates or vice versa. Generally, the registration needs many the point features or line features extracted automatically from images, then matching them and use the correspondence to calculate homography. Farin et al. [11] use a model of the arrangement of court lines for registration, similar in [8]. However, if not obvious lines in the fields or counts, it can not work well. Okuma et al. [7] compute the local displacements of image features using the Kanade-Lucas-Tomasi (KLT) tracker [12, 13] and determine local matches, but it needs predict the current camera motion based on the previous camera motion to reduce the amount of local features motion, if the camera moves rapidly and asymmetrically such as our application, the worse prediction can fail the KLT tracker.

Therefore, for a rapid camera, the registration face two big difficult problems: one is the detection of the more distinctive point features and matching them better. The other is how to reduce the accumulative registration error for a long image sequence.

3.1 Homography

A homography is a projective transformation represented as a nonsingular matrix 3×3 H. If x and x' are images of the same world point, belonging to a plane, they are related by a matrix H corresponding to that plane: $x' = Hx$, since the matrix H has 8 DOF, 4 point correspondences determine H. Obviously, with non-perfect data, more points should be used.

In general, the moving targets in the image can decrease the computational accuracy of homography since the good features on the moving ones are imperfect. RANSAC algorithm [14, 1] can pick out that worse features (namely outliers) easily, shown in Fig. 4. However, that can only assure the computational accuracy between two adjacent frames, not for a long video sequence. How to reduce the accumulative error is the key issue.

3.2 Selection of point features

The performance of point features extraction and matching determines the computational precision of the homography and the overall reliability of the registration algorithm. Two of the most popular feature are the Harris corner detector [15] followed by Sum of Squared Difference (SSD) matching, and the Kanade-Lucas-Tomasi (KLT) tracker [13, 12]. These methods can work well when the baseline is relatively small and the appearance of the features doesn't change too much across subsequences. Therefore, a more distinctive feature is desirable to suit for the matching of two wide baseline images (e.g. a large translation, scaling, or rotation between two frames) such as the current frame and its corresponding reference frame in our application.

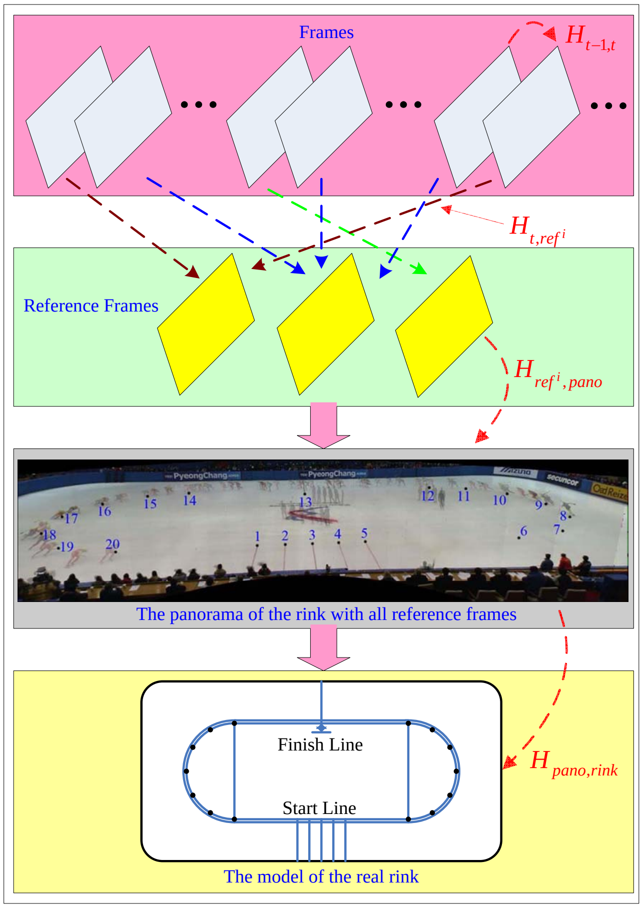


Fig. 2. The flowchart of the registration method for a long image sequence.

Lowe [16, 17] first proposes The Scale Invariant Feature Transformation (SIFT), where a feature's location and scale are determined by extrema of a DoG function in scale space, and its orientation by the dominant, local image gradient orientation. It is invariant to viewpoint changes, the large geometric transformation and change in illumination, and that has been applied in the areas of object recognition [18] and panorama stitching [19]. Therefore, it is superior to potentially deal with the problem of wide baseline matching.

3.3 Registration method for a long image sequence

In order to reduce the accumulation of errors produced from the set of frame-to-frame homography, many reference frames are calculated to construct a panoramic image [20, 19] as illustrated in Fig. 2 and each frame can be mapped to the most adjacent reference frames. The detailed model of the entire rink is shown in

Fig. 2, it includes precise measurements of geometrical features like start line, finish line and marking blocks, it can be obtained from ISU [21]. The corresponding points labeled from 1 to 20 are initialized by manually, the detailed algorithm is illustrated in Algorithm 1.

Algorithm 1 Registration method for a long image sequence

```
Input video sequence and perform following steps:
1. Compute homography $H_{t-1,t}$ between the frame at time $t-1$ and the frame at time $t$ with RANSAC algorithm [1]
2. Choose the frame in central angle of view of the whole rink as a base frame used to construct the panorama
3. Compute reference frames distributing on both sides of central frame at intervals
4. Generate the panorama of the rink with all reference frames and calculate homography $H_{ref^i,pano}$ transforming the $i^{th}$ reference frame to the panorama
5. Map the panorama to the rink in the world by selecting 20 corresponding points by manually shown in Fig. 2 and obtain homography $H_{pano,rink}$
6. Compute homography $H_{t,ref^i}$ mapping the frame at time $t$ to the corresponding reference frame
7. Obtain homography $H_{t,rink}$ mapping the frame at time $t$ to the rink in the world
   $H_{t,rink} = H_{pano,rink} \cdot H_{ref^i,pano} \cdot H_{t,ref^i}$
```

4 Experiments and Results

All of our tests were carried out on video recorded live by the authors in the World Short Track Speed Skating Championships.

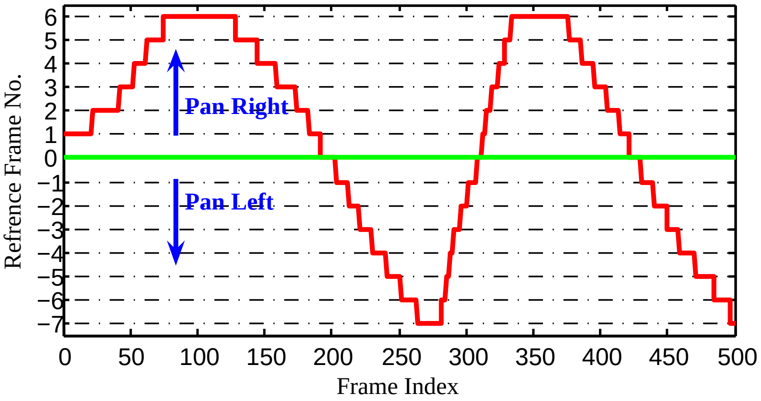
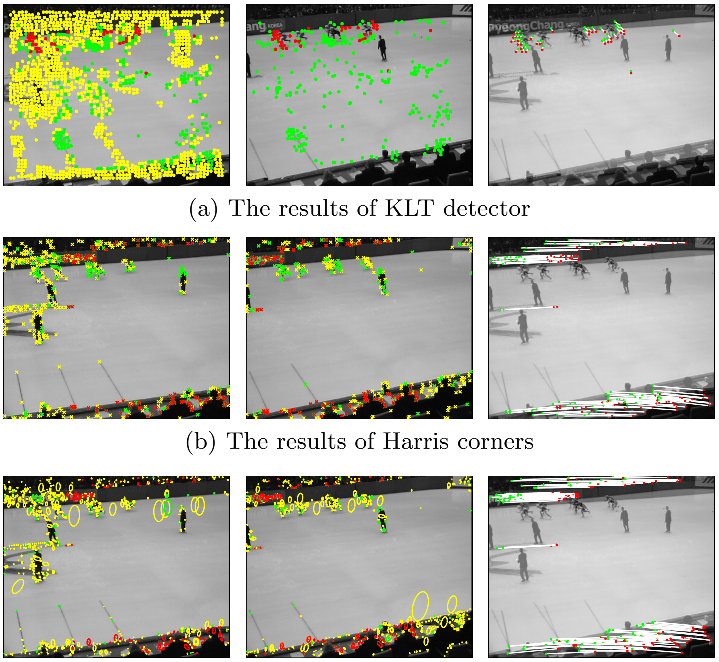


Fig. 3. The relation between each frame and corresponding reference frame.



(a) The results of KLT detector

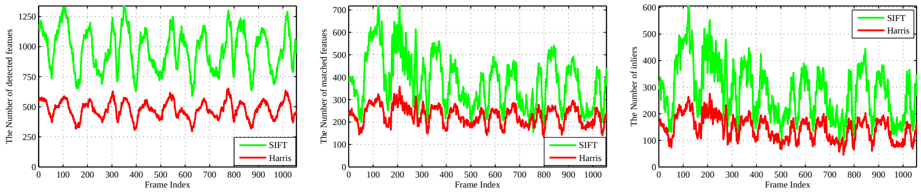
(b) The results of Harris corners

(c) The results of SIFT features

Fig. 4. The comparison of different features. The first column: all detected point features in current frame are shown, yellow points denote mismatch ones putatively by matching algorithm, green and red points represent outliers and inliers respectively, as a result of the homography estimation through a RANSAC procedure. The second column: the corresponding reference frame. The last column: the results are refined using RANSAC, red and green points are inliers of the current frame and reference frame respectively.

4.1 The relation between each frame and corresponding reference frame

The relation between each of 500 frames in a tested video and 14 reference frames is illustrated in Fig. 3, the specified base frame close to the center view can be selected by manually and the other 13 reference frames are calculated automatically. The blue label ”Pan Right” denotes that the camera pans from the center (reference frame No. is 0, namely the specified base frame) to the right. At frame $75 - 130$, the player skates on the right curve, the camera pans slowly, so the red line is wide, that means the corresponding reference frame of frame $75 - 130$ is No. 6 all along. At frame $130 - 210$, the skater moves through the straight from the right curve to the left curve, camera pans quickly, so the red lines become narrow. If the homography between frame 100 and 350 will be calculated, our method is $H_{100,350} = H_{ref\_6,350}H_{100,ref\_6}$, instead of $H_{100,350} =$



(a) All features (b) Matched features (c) Inliers

Fig. 5. SIFT VS Harris.

$H_{349,350} \cdots H_{t-1,t} \cdots H_{100,101}$, similar in [7]. As a result, the introduction of the reference frames can significantly reduce the accumulative error in theory.

4.2 The comparison of different features

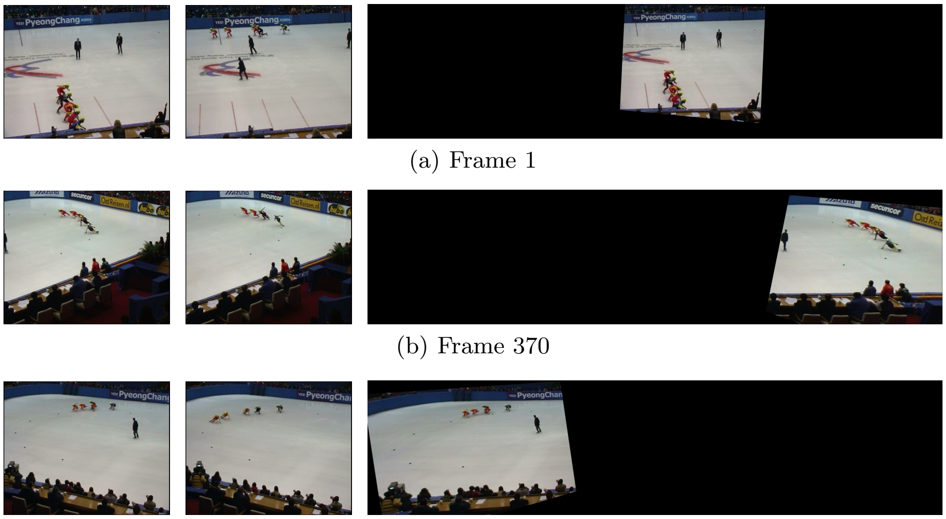
The comparison results of different features are illustrated in Fig. 4, where SIFT features are extracted from the current frame, total 980 (yellow, red and green), 149 inliers (red) and 103 outliers (green), at the same time, Harris: 473, 82 and 87, KLT: 1500, 39 and 259 (obviously wrong). That suggests SIFT and Harris corners are more suitable for a pair of wide baseline images (large translation) than the KLT detector. In Fig. 5, SIFT can be compared with Harris from the number of all detected features, matched features putatively and inliers on a test video over 1000 frames, SIFT can provide more reliable corresponding point features than Harris for our registration algorithm, it is helpful for improving the computational precision of the homography.

4.3 Analyse the precision of our system

There are 14 marker blocks on the rink and their spatial positions are prior known [21]. The imagery position of all visible markers in each frame are recorded by manually and transformed to the real world position by multiplying $H_{t,rink}$ that has been calculated by our system and compared with their ground truth value. The precision of our system is shown in Table 1 and Table 2, the frame number denotes the total statistical frame number of each marker. The mean error comes mainly from $H_{ref_i,rink}$ and H_{t,ref_i} determines the Gauss std. The results of automatic registration are shown in Fig. 6.

5 Conclusion

In this paper, we propose a robust approach of automatic registration for a long image sequence, that introduces the reference frames as a transition through which each frame can be mapped to the field model in order to reduce the error accumulation of the projection. Compared with Harris corners and KLT features,



(a) Frame 1

(b) Frame 370

(c) Frame 708

Fig. 6. The results of automatic registration. The first column shows the original image (720 × 576) and its corresponding reference frame is illustrated on the second column, the last column shows the registration result that mapping the original image to the model of the real rink.

Table 1. Registration error of 7 markers on the left curve (Units: m).

	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6	Marker 7
mean x	0.37	0.32	0.03	0.18	0.16	0.06	0.27
std x	0.04	0.04	0.06	0.05	0.04	0.04	0.03
mean y	0.35	0.28	0.13	0.10	0.21	0.21	0.07
std y	0.06	0.08	0.07	0.06	0.04	0.04	0.03
frame number	186	280	248	201	139	168	246

SIFT adopted by us has more advantages for two-frame wide baseline matching. Experiments show that the precision of our system is very high. Our system can be applied to various sports such as soccer, football, volleyball, basketball,hockey. With the development of hardware and software, it will become a online system for sports broadcast.

References

1. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2000)
2. : http://www.dartfish.com.
3. : http://www.simi.com.
4. : http://www.trakus.com.

Table 2. Registration error of 7 markers on the right curve (Units: m).

	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6	Marker 7
mean x	0.34	0.31	0.03	0.26	0.26	0.04	0.30
std x	0.05	0.06	0.04	0.04	0.04	0.03	0.02
mean y	0.28	0.15	0.04	0.01	0.11	0.19	0.16
std y	0.09	0.09	0.07	0.06	0.05	0.03	0.04
frame number	365	382	397	326	244	263	339

5. Junior, B.M., Anido, R.D.O.: Distributed real-time soccer tracking. In: ACM 2nd international workshop on VSSN. (2004) 97–103
6. Pers, J., Bon, M., Kovacic, S., Sibila, M., Dezman, B.: Observation and analysis of large-scale human motion. Human Movement Science **21**(2) (2002) 295–311
7. Okuma, K., Little, J.J., Lowe, D.G.: Automatic rectification of long image sequences. In: Asian Conference on Computer Vision (ACCV'04). (2004)
8. Hayet, J., Piater, J., Verly, J.: Robust incremental rectification of sport video sequences. In: British Machine Vision Conference (BMVC), London (2004)
9. Intille, S.S., Bobick, A.F.: Tracking using a local closed-world assumption: Tracking in the football domain. MIT Media Lab Perceptual Computing Group Technical Report 296 (1994)
10. Hayman, E., Murray, D.W.: The effect of translational misalignment when self-calibrating rotating and zooming cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(8) (2003) 1015–1020
11. Farin, D., Krabbe, S., de With, P., elsberg, W.: Robust camera calibration for sports videos using court models. In: Proc. SPIE Storage and Retrieval Methods and Applications for Multimedia. 80–91
12. Shi, J., Tomasi, C.: Good features to track. In: IEEE Conference onComputer Vision and Pattern Recognition (CVPR'94), Seattle (1994)
13. Tomasi, C., Kanade, T.: Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University (April 1991)
14. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6) (1981) 381–395
15. Harris, C., Stephens, M.: A combined corner and edge detector. In: Fourth Alvey Vision Conference, Manchester, UK (1988) 147–151
16. Lowe, D.G.: Object recognition from local scale-invariant features. In: International Conference on Computer Vision ICCV, Corfu. (1999) 1150–1157
17. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. **60**(2) (May 2004) 91–110
18. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: International Conference on Computer Vision ICCV. (2003) 1470–1477
19. Brown, M., Lowe, D.: Recognising panoramas. In: International Conference on Computer Vision ICCV. (2003) 1218–1225
20. Yeung, H.Y., szeliski, R.: Systems and experiment paper: construction of panoramic image mosaics with global and local alignment. International Journal of Computer Vision **36**(2) (2000) 101–130
21. : http://www.isu.org/.