ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.9, September 2006, pp.1899−1907
DOI: 10.1360/jos171899

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种整体的视频匹配方法*

柴登峰[1,2+], 彭群生[1]

[1](浙江大学 计算机辅助设计与图形学国家重点实验室,浙江 杭州 310027)

[2](浙江大学 空间信息技术研究所,浙江 杭州 310027)

A Global Approach for Video Matching

CHAI Deng-Feng[1,2+], PENG Qun-Sheng[1]

[1](State Key Laboratory of CAD & CG, Zhejiang University, Hangzhou 310027, China)

[2](Institute of Space Information Technique, Zhejiang University, Hangzhou 310027, China)

+ Corresponding author: Phn: +86-571-88206681, Fax: +86-571-88206680, E-mail: chaidf@cad.zju.edu.cn

Chai DF, Peng QS. A global approach for video matching. *Journal of Software*, 2006,17(9):1899−1907.
http://www.jos.org.cn/1000-9825/17/1899.htm

Abstract: This paper presents a new framework for spatiotemporal alignment of two video sequences. It proposes Intra-video and inter-video matching strategy for spatial alignment; modifies Dynamic Time Warping for temporal alignment. Intra-video matching tracks feature points and binds them together. Contextual inter-video matching uses track correspondences to provide initial feature correspondences for inter-video frame matching and updates track correspondences using frame-matching results. The proposed matching strategy makes best use of coherency of source videos and improves coherency of aligned video, stability and efficiency of alignment. The Modified Dynamic Time Warping establishes frame correspondences by minimizing global differences between them, keeps temporal order of frames, and handles nonlinear misalignment of videos. The proposed method can successfully align videos viewing different events recorded by independently moving cameras. Experimental results and comparison show that great improvements on stability and efficiency of video matching together with coherency of aligned video are reached.

Key words: video matching; video manipulation; video synthesis

摘 要: 给出一种视频时空配准的整体方法,提出一种视频内匹配与视频间匹配相结合的空间配准策略,改进动态时间扭曲方法以用于时间维的对齐.视频内匹配跟踪视频内各帧图像的特征点并记录其轨迹,视频间匹配配准不同视频的帧图像,使用轨迹对应提供图像配准所需的初始特征点对应,根据图像配准得到的特征点对应建立和更新轨迹对应.该匹配策略充分利用了视频的连贯性提高了匹配的稳定性和效率,同时提高了配准视频的连贯性.改进的动态时间扭曲方法通过极小化两段视频的整体距离建立视频之间的帧对应关系,保持视频内部各帧之间的时序关系并能处理非线性偏移.给出的方法能成功匹配独立运动相机在不同时刻拍摄关于同一

* Supported by the National Natural Science Foundation of China under Grant No.60273053 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.2002CB312101 (国家重点基础研究发展规划(973))

Received 2005-05-12; Accepted 2005-08-25

场景不同目标的视频.实验结果表明利用整体匹配方法,匹配的稳定性与效率和配准视频的连贯性都得到很大的提高.

关键词: 视频匹配;视频操作;视频合成
中图法分类号: TP391 文献标识码: A

Image alignment[1] has been studied extensively in the past decades. However, its potential applications are restricted by the limited information contained in the static images. Fortunately, video camera, which is capable of capturing image sequence and recording much additional information, is becoming more and more common in our everyday life. Alignment of video sequences would allow more applications, such as controversy settlement in football game[2]. Video alignment is becoming an active research area in both computer vision and graphics recently, but much interest has been put on videos viewing the same event.

For computer graphics, image and video manipulation is more attractive[3,4]. Sand and Teller showed that aligning two videos recorded at different times (viewing different events) allows a lot of potential applications[5]. But they didn't make full use of coherence of source videos. This impairs both stability and efficiency of video matching, and also impairs coherency of aligned video.

This paper discusses how to explore coherence of source videos to improve video matching. First, we propose intra-video matching stage to track feature points. Second, we introduce track correspondences to improve inter-video frame matching. The benefit of this matching strategy is that both stability and efficiency of alignment are improved, and the spatial coherency of aligned video is also improved. Third, we apply and modify Dynamic Time Warping[6] for temporal alignment. It keeps temporal order of frames and handles nonlinear misalignment, therefore improves temporal coherency of aligned video. All these improvements are important for graphics applications.

The remainder of this paper is organized as follows. First, related work on video alignment is remarked in Section 1. In Section 2, we formulate the video matching problem dealt with in this paper. In Section 3, we propose intra-video and inter-video matching strategy for spatial alignment. In Section 4, we modify Dynamic Time Warping for temporal alignment. The framework for spatiotemporal video alignment is presented in Section 5. In Section 6, we show some experimental results and make comparisons with the method presented in Ref.[5]. At last, we draw conclusion in Section 7.

1 Related Work

Image matching is an active research area in computer vision. Image registration methods[1] try to find a parametric transform mapping one image to the other image. As the transform is applied to the entire image, it works well only for images captured by rotating camera or images of planar scene. Optical flow methods[7] try to find for every pixel in one image the offset vector to the corresponding pixel on the other image. Because different pixels may have different offset vectors, they allow small change of viewpoint and variance of depth of the scene from the viewpoint. Stereo matching methods[8] try to establish dense pixel correspondences between two images under the epipolar geometry constraint; the search for correspondence is restricted to 1D line instead of 2D plane. These methods are designed for wide-baseline stereo images and suitable for 3D shape reconstruction.

Recently, video alignment is becoming an active research topic. Caspi and Irani laid an assumption that the video sequences are related by a homography and a constant time offset, and then solved the unknown parameters either by feature-less[9] or feature-based (space-time trajectories)[10] method. Cameras need to be static or connected together rigidly to insure that the whole sequences are related by a single homography and a constant time offset.

Tutelaars and Van Gool[11] proposed a method for temporally aligning video sequences that allow cameras to move independently and to have constant time offset. It needs a configuration of 5 moving points matched and tracked through both sequences. Carceroni etc.[12] introduced *N*-dimension timeline to align *N* sequences and adopted linear function to handle different frame rates. The cameras can be placed at distinct viewpoints but need to be static. Rao etc.[13] proposed a method for temporal alignment of videos of human activities. They allow cameras to be placed at distinct viewpoints and have nonlinear time offset. The method needs user selection of trackable feature points. All these works deal with videos viewing the same event; besides, they deal only with temporal alignment.

Sand and Teller presented a method for spatiotemporal alignment of videos captured at different time by moving cameras[5]. The assumption is that video sequences must have spatially similar motions (nearly the same trajectory through space) and images contain enough texture. Cameras can move in space, image appearance can be different (change of lighting, exposure and object), but coherence of videos is not well utilized.

2 Problem Statement

The goal of our work is to align videos reordered under the conditions: first, videos are recorded at different times, the background is the same but foreground may be different (or the scene is the same but event may be different), second, cameras move independently but follow nearly the same trajectory through space. Because the cameras are moving independently, the temporal misalignment may be nonlinear function, and, the spatial mapping functions may vary with time (frames). These distinguish from other works.

Stated formally, given primary video V_1 and secondary video V_2, find a *temporal mapping function* $T(i),i=1,\ldots,N$, (N is the total number of frames of V_1) mapping each frame of V_1 to its corresponding frame in V_2. And, for each corresponding frames F_i and $F_{T(i)}$, find a *spatial mapping function* $(x'y')=S_i(x,y),x=1,\ldots,C,y=1,\ldots,R$ (R and C are total number of rows and columns of the video frame respectively) mapping each pixel $(x,y)$of F_i to its corresponding pixel $(x'y')$ in $F_{T(i)}$.

Sand and Teller[5] developed a robust image alignment method to estimate the spatial mapping function. First, it detects feature points in images and establishes initial feature correspondences based on pixel consistency, then, it combines Locally Weighted Regression[14] and EM algorithm[15] to find good correspondences. At last, it uses the good correspondences to interpolate and extrapolate the spatial mapping function. All frame matching is carried out independently, and the initial correspondences are established according to pixel consistency. We refer this as *independent frame matching* strategy. When images to be aligned have different appearances, the initial correspondences may contain many outliers inevitably, then, the iterative algorithm may be ineffective or even fail. In Section 4, we will propose the *intra-video and inter-video matching* strategy to improve it.

They also use good correspondences to evaluate frame similarity and develop a temporal alignment algorithm based on the frame similarity measure: To find $T(i)$, it gets an initial guess from the previous frames, evaluates the similarity between F_i of V_1 and frames of V_2 near the initial position and then applies the "local winner takes all" schema to select the most similar frame. This method may lead to *temporal incoherency*:

- Inverse frame: The latter frame of V_2 corresponds to the former frame of V_1, while the former one corresponds to the latter one.
- Freezing frame: A number of adjacent frames of V_1 correspond to the same frame in V_2.
- Jumping frame: The corresponding frames for consecutive frames of V_1 are not consecutive and have a large gap.

As for spatiotemporal manipulation, the objects in videos are moving. The temporal incoherency would make

aligned video useless. In Section 4, we will modify Dynamic Time Warping[6] to deal with these problems.

3 Spatial Alignment

When videos are recorded at different times, they may have different appearances. Then, the image alignment method presented in Ref.[5] may be ineffective or even fail to align frames across two videos. But the successive frames within one video have nearly the same appearance; the initial correspondences with few outliers can be obtained easily. In this case, the image alignment method is robust and effective.

We propose *intra-video matching* and *inter-video matching* strategy for spatial alignment. As shown in Fig.1, intra-video matching matches successive frames *within* each video and inter-video matching matches frames *across* the two videos. Intra-video matching matches successive frames to track feature points corresponding to the same scene point and binds them together as a *track*. Because each feature point belongs to one track, feature correspondence implies track correspondence, and vice versa. The feature correspondences can be used to establish track correspondences and track correspondences can be used to provide initial feature correspondences. We will present the details in Subsection 3.2.

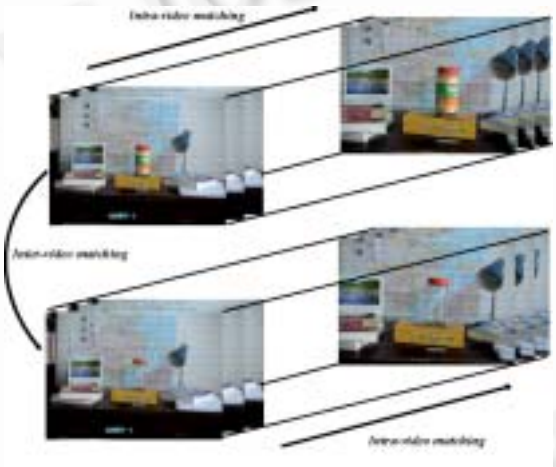

Fig.1 Intra-Video and inter-video matching

3.1 Intra-Video matching

Intra-video matching matches intra-video frames. First, it selects feature points in each frame using Harris corner detector[16], then establishes feature correspondences for all consecutive frames.

If the baseline between successive frames is short enough when compared with the distance between the scene and the viewpoint, we can use a homography to relate the corresponding points, as shown in Eq.(1), otherwise, we can use a fundamental matrix to relate them, as shown in Eq.(2).

$$x'=H_x \tag{1}$$

$$x'F_x=0 \tag{2}$$

where x and x' is the homogeneous coordinates of the corresponding points in two frames. H and F are 3-by-3 matrixes denoting the homography and fundamental matrix.

Then, a method like that of Ref.[17] is adopted to find good correspondences between two frames. First, it obtains some putative correspondences according to pixel consistency. The “Winner takes all” scheme is applied to guarantee that one point in one image matches only one point in the other image. After that, RANSAC algorithm[20]

is applied to estimate the homography (or fundamental matrix) and find inliers. At last, it uses the estimated homography to find more correspondences and to refine the homograpy in turn. This is iterated until the number of correspondences is stable.

3.2 Inter-Video matching

Inter-video matching matches inter-video frames. Our frame matching is not independent but contextual. It is benefited from our introduction of the *track correspondences*. Track correspondence is a correspondence of the two tracks in two videos respectively; they correspond to the same scene point. When inter-video frames are matched successfully, the corresponding feature points can establish (the first frame matching) or update (the rest of frame matching) the track correspondences:

If frame F_1 from V_1 and F_2 from V_2 are matched successfully and point P_1 in F_1 corresponds to P_2 in F_2, P_1 and P_2 belong to tracks T_1 and T_2 respectively, then, set T_1 correspond to T_2. Applying it to all feature points in F_1 can establish the track correspondences.

Essentially, we adopt the frame matching method of ref.[5], but use different strategy to provide the initial feature correspondences. For the first frame matching, it establishes the initial correspondences according to pixel consistency. For the rest of frame matching, it establishes the initial correspondences according to the track correspondences:

Given F_1 from V_1 and F_2 from V_2 to be matched, P_1 in F_1 and P_2 in F_2 belong to T_1 and T_2 respectively, if T_1 correspond to T_2, then, set P_1 corresponds to P_2. Applying it to all feature points in F_1 will establish the initial feature correspondences.

The key to the iterative method is that the initial correspondences must be good enough to ensure the iteration converging to the true solution. Even when convergence is not a problem, better initial correspondences are always preferable because they can speed up the matching. The initial feature correspondences provided by track correspondences are results of the previous frame matching, and they are much better than that provided by pixel consistency comparison. Besides, initial correspondences provided by track correspondences are stable across successive frames, therefore there exists so strong consistency among them that frame matching is stable and spatial coherency of the aligned video is improved. These are verified by our experimental results.

4 Temporal Alignment

As shown in Section 2, the “local winner takes all” schema for temporal alignment may lead to temporal incoherency. Dynamic Time Warping is widely used for temporal signal alignment[6], it aligns signals by minimizing a global distance between the signals, and it handles nonlinear misalignment while keeping temporal order of the signals. It was applied to video alignment by Rao etc.[13].

Figure 2 shows how DTW works. Each cell $D(i,j)$ of the *distance matrix D* stores the distance between sample i of signal S_1 and sample j of signal S_2, and DTW searches the best path from the cell $D(0,0)$ to the cell $D(N,N)$. The best path is the one with the least global distance, which is the sum of cells on the path. Each cell $D(i,j)$ on the best path indicates that sample i of S_1 corresponds to sample j of S_2. Although there are exponentially many such paths, the best one can be found in only quadratic time by the use of dynamic programming[6]. It uses matrix E to record the least global distance between signals aligned up to current

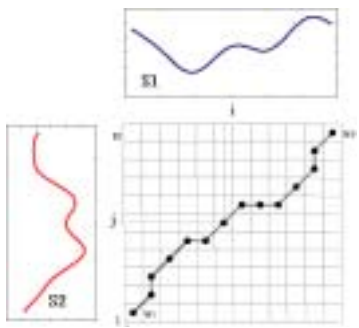

Fig.2 Dynamic time warping

sample, i.e. $E(i,j)$ stores the distance between signals aligned up to samples i and j of S_1 and S_2 respectively.

We modify DTW as follows for video alignment, as illustrated in Fig.3.

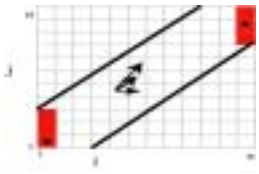

Fig.3 Modified DTW

First, replace S_1 and S_2 with V_1 and V_2, then replace sample points with video frames, measure the distance between frames using frame similarity measure presented in Ref.[5] based on the point correspondences result from inter-video frame matching. By adopting Eq.(3) to fill in D, it fills cells far from the diagonal (as illustrated by the two diagonals) with ∞ because it is impossible to make such frames correspond to each other. This also speeds up computation.

$$D(i,j)=\begin{cases} dist(i,j) & \text{if } |i-j|\le beamwidth \\ \infty & \text{if } |i-j|> beamwidth \end{cases} \tag{3}$$

where $dist(i,j)$ is the frame similarity between frames i and j, $beamwidth$ is specified in Section 6.

Second, adopt Eq.(4) to fill in E, and this constrains the cells to move in such a manner (as illustrated by the three arrows): it moves 1 step forward each time in the horizontal direction corresponding to V_1, meantime, it can move 0,1 and 2 steps each time in the vertical direction corresponding to V_2.

$$E(i,j)=D(i,j)+\min(E(i-1,j),E(i-1,j-1),E(i-1,j-2)) \tag{4}$$

Third, search the best path from one of the left-bottom cells (red ones in the figure) to one of the right-up cells. Select the minimal one of the right-up cells of E and trace back to one of the left-bottom cells. This allows two videos to have different frames and the first corresponding frame to shift in some degree without any extra cost.

After the best path is found, if it passes $D(i,j)$, then, set the temporal mapping function $T(i)=j$.

As it can be seen in Figs.3 and 6, MDTW selects cells by minimizing the global distance, constraining the path to move smoothly, and keeping the temporal order, this distinguishes from the "local winner takes all", which selects cells only according to the local distance. In some sense, it maintains the temporal constraint at the cost of spatial similarity. If the videos are recorded by hand-held cameras, they contain revisited scene inevitably. If the revisited frames don't exceed *beamwidth*, MDTW can align videos successfully at the cost of spatial similarity. But, MDTW will fail to handle large revisited frames. Fortunately, videos used for temporal manipulation contain nearly the same revisited scene, and then it can be treated as a non-revisited case. If the revisited scene are different for two videos and there are no moving object in the secondary video, we recommend to use the "local winner takes all" instead.

5 Framework of Spatiotemporal Video Alignment

In this section, we bring Sections 3 and 4 together to build up a framework for spatiotemporal alignment of video sequences.

(1) Feature detection: Apply Harris corner detector to detect feature points in each frame of each video.

(2) Intra-Video tracking: Carry out intra-video matching to track feature points for each video.

(3) Temporal alignment: Carry out MDTW to find temporal mapping function as shown in Section 4.

(4) Spatial alignment: Carry out inter-video matching as shown in Subsection 3.2 to spatially align each corresponding frames found in step 3, establish point correspondences (KLT optimizer[19,20] is applied to refine alignment to sub-pixel precision), interpolate and extrapolate the spatial mapping functions.

6 Experimental Results

We applied our method to a variety of videos recorded by moving cameras, some matching results and comparison with that of Ref.[5] are presented in this section.

To compare feature points, we use 200 by 200 window centered at the point as proximity constraint, use 15 by 15 window to compare feature points, use the average distance to the nearest 50 points to set the adaptive kernel width for Locally Weighted Regression, and set σ_{pixed}=2 and σ_{motion}=2. The *beamwidth* for DTW is set to 10.

Figure 4 shows the result of intra-video matching (Eq.(1) is used because the baseline is short enough here) of a 200-frames video, each vertical line denotes one track, and the upper extent of the line is the frame index at which the point is fist detected. The computation time of the intra-video matching is 7.77 seconds.

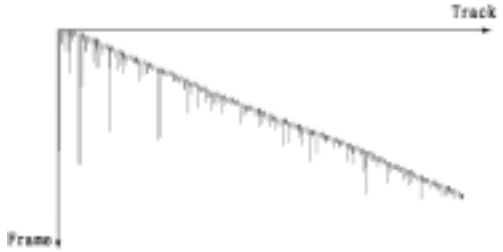

Fig.4 Result of intra-video matching

Figure 5 compares the independent frame matching strategy with the intra-video and inter-video matching strategy. The first two rows show frames from V_1 and V_2 (the selected frames are 0,1,99,199). The next two rows show initial correspondences established using the two strategies. The 3rd row corresponds to the independent frame matching strategy, and they are results of the pixel consistency comparison. Obviously, there are many outliers indicated by the blue lines. The 4th row corresponds to inter-video and intra-video matching strategy. As shown, the first frame has the same correspondences as that of the 3rd row, but the second and the following frames have much better initial correspondences. As a result, the average number of iteration is 2 while that of the independent frame matching is 5. The computation time is reduced from 0.059 to 0.021 second per frame. When videos have 200 frames, the total time of distance matrix computation is 243.29 seconds for independent frame matching and 84.90 seconds for intra-video and inter-video matching strategy. When KLT optimizer is applied for spatial alignment, total time is reduced from 1239.90 to 359.97 seconds. The 5th and 6th rows show good correspondences found in the two strategies. There are still some outliers in the 5th row (independent frame matching), but there are few outliers in the 6th row (intra-video and inter-video matching). The last two rows show the aligned images in the two strategies

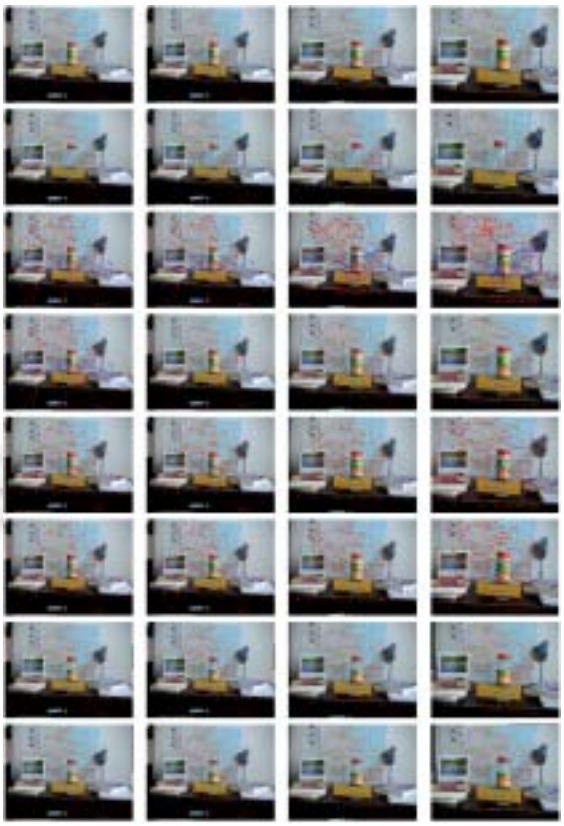

Fig.5 Comparison of spatial alignment using two different matching strategies: circles indicate feature points and lines from points indicate offset vectors to their corresponding points

and, as expected, the last row (intra-video and inter-video matching) is better. The stability and spatial coherency of the aligned videos are also improved by our method.

Figure 6 shows the temporal alignment results, when the "local winner takes all" is applied, there are temporal incoherencies, but they are removed by MDTW. Figure 7 shows two examples of temporal incoherency, the first two rows show the frames from V_1 and V_2, and the last row shows the aligned images. The frames 25,...,33 of V_1 have the same corresponding frame 26 in the V_2, and it is a case of freezing frame and illustrated in the left two columns. The corresponding frames of frames 62 and 63 in V_1 are 62 and 71 respectively; it is a case of jumping frame and illustrated in the right two columns.

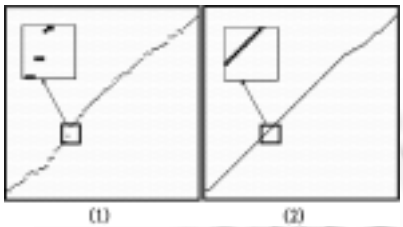

Fig.6 Frame correspondences found by local winner takes all (1) and MDTW (2)

Fig.7 Temporal incoherency

7 Conclusion

This paper discusses how to capitalize upon coherence in source video to improve video matching.

We propose intra-video and inter-video matching strategy for spatial alignment. Intra-video tracks points corresponding to the same scene point, and contextual inter-video matching makes full use of the consistency within each track. It improves both stability and efficiency of alignment, and improves spatial coherency of aligned video also.

We also modify DTW for temporal alignment. MDTW keeps temporal order of the frames and handles nonlinear misalignment, it establishes frame correspondences by minimizing global distance between two videos, and therefore temporal coherency of the aligned video is improved at the cost of spatial similarity. MDTW can handle small revisited frames (it exists inevitably in the hand-held camera case). If too many frames are revisited in one video with respective to the other, or, the relative speed of one camera with respect to the other is large, MDTW may fail to align the videos. Fortunately, it occurs rarely in the temporal manipulation applications. We plan to develop an algorithm to cut videos into segments and align videos segment by segment in the future.

Besides, we still assume that cameras follow the similar trajectories as that of ref.[5], but we believe that this limitation may be partially overcome by using intra-video and inter-video matching strategy. We plan to overcome this limitation by combining intra-video matching and wide-baseline stereo matching technique in the future.

References:

[1] Zitová B, Flusser J. Image registration methods: A survey. Image and Vision Computing, 2003,21(3):997−1000.

[2] Reid I, Zisserman A. Goal-Directed video metrology. In: Proc. of the European Conf. on Computer Vision'96. 1996. 647−658.

[3] Chuang YY, Agarwala A, Curless B, Salesin DH, Szeliski R. Video matting of complex scenes. In: Proc. of the SIGGRAPH 2002. 2002. 243−248.

[4] Agarwala A, Dontcheva M, Agarwala M, Drucker S, Colburn A, Curless B, Salesin D, Cohen M. Interactive digital photomontage. In: Proc. of the SIGGRAPH 2004. 2004. 294−302.

[5] Sand P, Teller S. Video matching. In: Proc. of the SIGGRAPH 2004. 2004. 592−599.

[6] Darrell T, Essa I, Pentland A. Task-Specific gesture analysis in real-time using interpolated views. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1995,18(2):1236−1242.

[7] Beauchemin SS, Barron JL. The computation of optical flow. ACM Computing Surveys, 1995,27(3):433−467.

[8] Scharstein D, Richard S. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int'l Journal of Computer Vision, 2002,47(3):7−42.

[9] Caspi Y, Irani M. A step towards sequence to sequence alignment. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition 2000. 2000. 682−689.

[10] Caspi Y, Irani M. Feature-Based sequence-to-sequence matching. In: Proc. of the Vision and Modeling of Dynamic Scene Workshop with European Conf. on Computer Vision 2002. 2002.

[11] Tuytelaars T, Van GL. Synchronizing video sequences. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition 2004. 2004.

[12] Carceroni RL, Pádua FLC, Santos GAMR, Kutulakos KN. Linear sequence-to-sequence alignment. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition 2004. 2004.

[13] Rao C, Gritai A, Shah M. View-Invariant alignment and matching of video sequences. In: Proc. of the IEEE Int'l Conf. on Computer Vision 2003. 2003. 939−945.

[14] Atkeson CG, Moore AW, Schaal S. Locally weighted learning. Artificial Intelligence Review, 1997,11(1):11−73.

[15] Duda RO, Hart PE, Stork DG. Pattern Classification. Beijing: China Machine Press, 2003.

[16] Harris CJ, Stephens M. A combined corner and edge detector. In: Proc. of the 4th Alvey Vision Conf.'88. 1988. 147−151.

[17] Hartley R, Zisserman A. Multiple View Geometry in Computer Vision. Cambridge: Cambridge University Press, 2000.

[18] Fischler MA, Bolles RC. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 1981,24(6):381−395.

[19] Lucas B, Kanade T. An iterative image registration technique with an application to stereo vision. In: Proc. of the Int'l. Joint Conf. Artificial Intelligence'81. 1981. 674−679.

[20] Shi J, Tomasi C. Good features to track. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition'94. 1994. 593−600.

CHAI Deng-Feng was born in 1974. He is a Ph.D. candidate at the State Key Laboratory of CAD&CG, Zhejiang University, He also serves as an assistant professor at Institute of Space and Information, Zhejiang University. His current research areas are vision based graphics and photogrammetry.

PENG Qun-Sheng was born in 1947. He is a professor and doctoral supervisor at State Key Laboratory of CAD&CG, Zhejiang University and a senior CCF member. His research areas are realistic image synthesis, computer animation, scientific data visualization, virtual reality, bio-molecule modeling.