Video Anatomy: Cutting Video Volume for Profile

Hongyuan Cai
Department of Computer Science, IUPUI
723 W. Michigan St. SL280, Indianapolis, IN46202
hocai@cs.iupui.edu

Jiang Yu Zheng
Department of Computer Science, IUPUI
723 W. Michigan St. SL280Indianapolis, IN46202
jzheng@cs.iupui.edu

ABSTRACT

This work takes a new aspect view to profile a video volume to a video track as a digest for video preview. Our projected video profile contains both spatial and temporal information inclusively in a 2D image scroll that is continuous, compact, scalable, and indexing to each frame. To profile various types of video clips, we investigate the global flow field under all camera actions, and propose a uniformed scheme that uses a sampling line to cut the video volume across the major optical flow field. The resulting profile obtains an intrinsic scene space less influenced by the camera actions, and can be displayed in a video track to guide the access to video frames, and facilitate video browsing, editing, and retrieval.

Categories and Subject Descriptors

H.2.4 [**Database Management**] Database Applications – *Multimedia databases.* I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis – *motion, time-varying imagery.*

General Terms

Algorithms, Measurement, Design, Experimentation.

Keywords

Video summary, video indexing, profile of video, camera motion, spatial-temporal slice, optical flow

1. INTRODUCTION

Current video indexing uses key frames for video clips/shots and their collection as story boards or scenes [1]. However, a key frame picks only a moment of the video clip. To index to each individual frame in a video summary, this work creates a 2D profile from the raw video data, which is an image belt containing one axis as the timeline and the other from a space dimension in the video. It is a novel aspect view of video that can index to each frame for video editing, and provides a view of entire scene space that a video captures for browsing. The created time belt can be presented in a video track in video editing and production, displayed in webpage for browsing, and used for video retrieval. Figure 1 displays a possible cutting for such a profile in the video volume.

To achieve above goals, this paper designs the cutting in the video volume to yield a planar or curved slice to reveal the video content in the video volume. It is implemented by sweeping a sampling line across the video volume. Our criteria to cut a profile from video are to include all the stable background space a video clip captures, to show meaningful shapes and identities of objects, and to reduce the distortion of target scenes in the profiles that obey different scene projections from the normal perspective projection. We analyze the typical camera works (motion styles), its underlying kinematics, and the generated optical flow in the video to ensure that our designed cutting and slicing strategy work for all types of video sequences. The significance to generate such a profile of video lies in its (i) compact size, (ii) reflecting temporal information, (iii) preserving shape to some extent, (iv) embracing static background and dynamic foreground, and (v) robustness in processing. Because dynamic video frames in a clip have overlaps on scenes, the reduction of redundant pixels in the 2D video profile becomes possible for video summarization. Our slicing of the video volume is designed to cut through every frame so that the continuous profile indexes to frames. The spatial information such as static environment and dynamic targets (objects, people, etc.) in the video is also visualized in the profile, although some deformation and changes in spatial ordering are brought in.

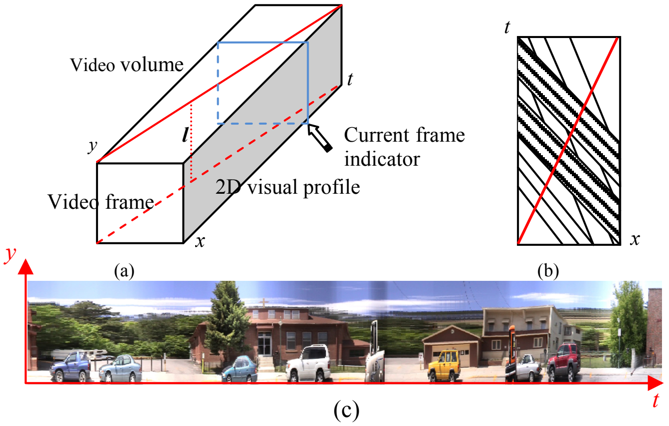


Figure 1. Video volume and a possible cutting of diagonal slice across the major flow in a video clip. (a) Video volume and a cutting slice, (b) an EPI showing flow traces of scenes, (c) profile of a video clip from a sideway translating camera.

The related works include *spatial* and *temporal video indexing*. The spatial indexing presents key frames [1] and its extension in a larger spatial domain generated by stitching camera panning views [2] and overlapping human actions in the same spatial domain [3-6]. These montage-based methods selectively stitch time-different regions into a video summary image. The drawbacks are as follows. (1) The camera motion is only limited to static and pan cases because no motion parallax is allowed (or single-depth background is required); (2) The extended key frame becomes cluttered if the video clip lasts for long or targets have a high complexity in the video; (3) The generated summary lacks of temporal order; (4) The matching of background and

*Area Chair: Lei Chen.

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

segmentation of foreground are not robust for complex and dynamic scenes. In contrast, the *temporal indexing,* though under the controlled camera motion, has good results on camera rotation and translation by fixed [7,8] and dynamic slit scanning [9]. The shortcomings are (1) low temporal resolution for a short clip with fast motion; (2) views with a different projection from perspective projection; (3) deshaking of video or post smoothing [10,11,12].

Our proposed video summary in the profile is easy to be embedded into video software to enhance video editing, retrieval, analysis, and visualization in general. Compared with the related works, our video profile is a spatial-temporal slice to overcome the problems of related methods in the resolution, camera motion type, and robustness. It works for video clips from all kinds of camera works (operations). Our slice cutting strategy avoids image matching, flow segmentation, and other complex procedures in order to achieve the robustness. The computation involved is the global flow detection in a video clip. Moreover, the created video summary keeps the temporal order of video scalable in time, and has a higher resolution than spatial or temporal indexing method. Figure 1 gives an example of such profiles from a linear camera motion. As an initial result, our profiling method has been working on simple camera actions such as zoom, pan, and translation [13]. This paper focuses on building a uniformed framework for the video volume cutting on general camera alignments and motions that are the combination of simple motions. We analyze the intrinsic camera works (operations) and categorize video clips, and then examine the flow components of static background caused by the camera motion and the dynamic human/target/object motion against the background.

2. CAMERA MOTION AND IMAGE FLOW

A camera can be static or undergo motions such as zoom, rotation, translation, and their combinations. Practically, a video can be categorized by its camera works (operations) such as pan/tilt rotation, rail/vehicle based movement (camera facing forward/ backward/sideways), focusing and around-object motion (composed of simultaneous translation and rotation), forward moving or zooming, and so on. In addition, dynamic scenes in a static field of view (FOV) may further reveal a variety of motions as directional motion (e.g., marathon crowds) or diversified motion (e.g., random walking in a shopping mall). In general, we can describe the camera kinematics [13], which yields typical camera works generating distinct optical flows that can be classified as diversified flow or directional flow. This categorization helps us design a general cutting strategy to obtain the profile of video clips.

Now, let us examine some common properties of the categorized motions. For a video clip/shot with directional flow generated from a smooth camera ego-motion or a directional movement of target crowds, we can specify a *major flow* component in FOV (Figure 2). Moreover, we consider the flow component orthogonal to the major flow as the *minor flow*, and other flow components such as instantaneous rotation (camera roll) and shaking as the *unstable flow* component. Denoting a video clip by $\boldsymbol{C}$, and the optical flow vector by $\boldsymbol{u}(x,y,t)$ in the clip, the major flow $\boldsymbol{M}\in R^3$ in the video volume is computed as

$$M = \frac{1}{N}\sum_{x,y,t\in C} u(x,y,t) \qquad (1)$$

where N is the number of points in the clip. It shows direction if the scenes have a global motion or the camera motion is smooth.

The minor flow component $\boldsymbol{m}$ orthogonal to $\boldsymbol{M}$ is from particular styles of camera motion, hand shaking, unstable walking, and vehicle waving during the video capture. Its effect is visible in our profile as tilt changes during panning, translation, and zoom. Minor flow can be kept in the profile to reflect the dynamics of the camera, or can be removed by a video deshaking algorithm before and after profiling [10,11,12]. The *unstable motion*, $\boldsymbol{r}$, is from the instantaneous camera rolling, as well as the motion of dynamic targets against the background (BG) including non-rigid articulate motions and the deformation of human body and face, fire, smoke, water, etc. This type of motion is not dominant in the video clip for its smaller regions than background. Otherwise, it will be treated as major flow. The minor and unstable motion will not be used to determine the slice cutting. Although the unstable flow can be deshaked to some extent, we avoid cosmetic rectification since recording true motion in the profile is useful in clip editing.

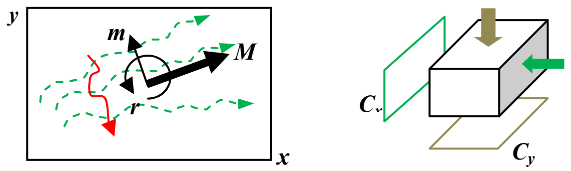


Figure 2. Major flow *M*, minor flow *m* and unstable flow including foreground motion (red) in video to be accumulated to condensed images.

We use *condensed image* used in vehicle video [10,11] to understand the distinct major flow in a video clip efficiently. In such slices, the motion is displayed as traces. Condensed images here are collected by averaging the intensities along the horizontal and vertical directions in the video. As an example in Figure 3, the major flow vector is close to horizontal direction. The long or high contrast features F_v perpendicular to $\boldsymbol{M}$ (i.e., in $\boldsymbol{m}$ direction) in the frames show their clear traces in one condensed image, while those features F_p parallel to $\boldsymbol{M}$ are blurred in it. Similarly, in the second condensed image, the condensing poses the stationary blur [14] on feature F_v but keeps F_p features sharp, and shows minor motion on them at the same time.

If the major flow is non-parallel to either x or y axes ($\boldsymbol{M}$ in a slanted direction), both condensed images contain motion and shape information such as traces and shapes with stationary blur. Depending on the dominant information that the condensed images contain, we refer to one as *motion-orientated* and the other as *shape-orientated*, respectively. One displays more traces from the linear features in the video, and the other shows more blurred shape than traces. These properties are used in the profile cutting next.

3. CUT VIDEO VOLUME FOR PROFILE

This section explores a general slice cutting framework for the profiles of video clips from various simple or combined motions.

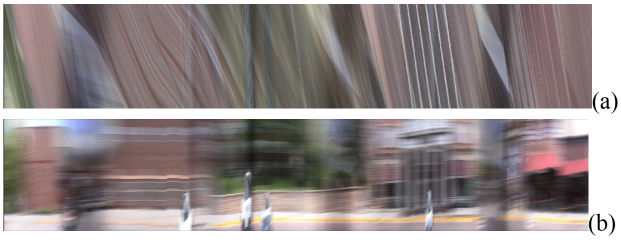


Figure 3. Two condensed images from a camera translating sideways. One is motion-oriented and the other is shape-orientated obtained by condensing intensities vertically and horizontally in the frames respectively. The time axes are horizontal.

Different from those methods of stitching regions from key frames, this work sets a sampling line $\boldsymbol{l}$ in the video frame to scan the major flow from one side to another across the frame. The ideal setting of $\boldsymbol{l}$ is to minimize $|\boldsymbol{l}\cdot\boldsymbol{M}|$ in the frame for cutting a sharp and complete shape (If $\boldsymbol{l}\perp\boldsymbol{M}$, then $|\boldsymbol{l}\cdot\boldsymbol{M}|=0$). At the same time, $\boldsymbol{l}$ should align with a set of structure lines in the frames, e.g., horizon, building rims, wall, etc., in order to preserve the scene structure in the generated profile. In real implementation, most clips are right side up aligning with structural lines in the 3D space, i.e., the video frame (axis or frame boundary) is aligned with structure lines in the scene. This is a reasonable condition for most video clips because of the gravity and the standing pose in taking and watching video. Now, we propose a qualitative cutting method for a general camera motion in taking a video clip.

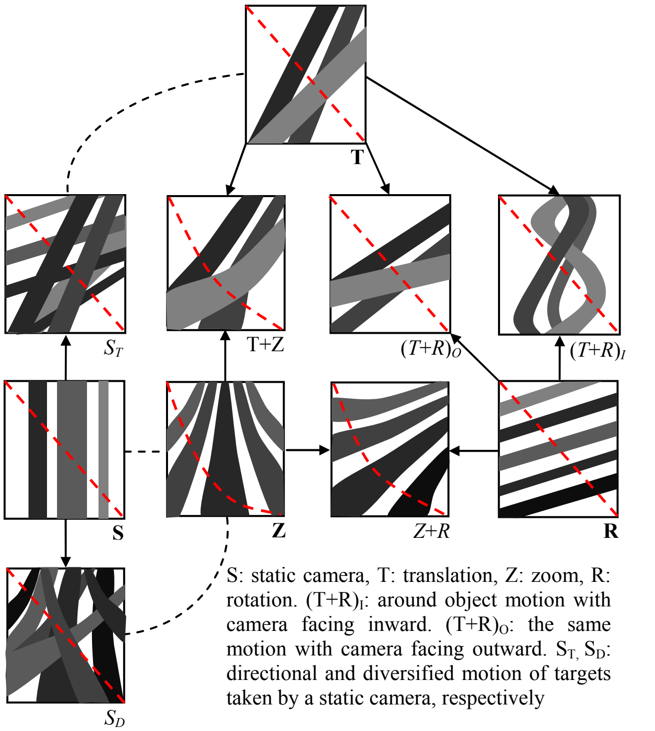
1. Alignment of cutting slice: sampling line $\boldsymbol{l}$ for the slice will be parallel to one of the frame edges in either horizontal ($\boldsymbol{x}$ axis) or vertical direction ($\boldsymbol{y}$ axis), which results in a profile maximally compatible to the video orientation. We examine if the major flow vector is closer to vertical direction than horizontal direction to determine the line orientation in the frame. A vertical line $\boldsymbol{l}$ is selected if $|\boldsymbol{M}\cdot\boldsymbol{x}|\geq|\boldsymbol{M}\cdot\boldsymbol{y}|$, and a horizontal line is selected otherwise.

2. Diagonal Cutting: A slice will be cut with the evolving line $\boldsymbol{l}$ diagonally across the video volume for more scenes. Which direction to cut depends on the major flow, i.e., the slice should cut across the motion traces in the volume rather than aligning with them so as to include more scenes sharply in the slice. The moving direction of $\boldsymbol{l}$ is opposite to the major flow direction in the frame, and thus is most likely consistent to the camera moving direction.

3. Trajectory bending: The bending of the cutting slice (the speed change of scanning) is determined according to the zoom factor or the divergence of the flow. We will bend the diagonal slice towards the enlarged frame in the zoom clip.

We explain this algorithm using the traces in the condensed image in Figure 4, which displays the flow characteristics of general camera motions. For the simplicity, we assume here the major flow vector is horizontal, i.e., the camera motion is horizontal. From the simple camera motion such as zoom, rotation, and translation (abbreviated to Z, R, and T in bold) and their corresponding slice cutting [13], Figure 4 shows all the possible motion combinations. For example, a camera can zoom during translation (T+Z). A translating camera with its optical axis non-orthogonal to the path is also a case of T+Z. On a circular path around a center (T+R), the camera can face inside as $(T+R)_I$ to focus on a target [15], where the background generates the major flow. If the camera faces outward, i.e., $(T+R)_O$, which is usually obtained with a crane arm, the generated flow is more similar to translation with motion parallax [16]. Moreover, cases Z and S_D, as well as cases T and S_T in Figure 4 have inherently the same types of flow by looking at their flow traces.

For all composite camera motions above (in between simple motions Z, R, T), their traces and slice cutting trajectories in our algorithm are depicted. According to the additive property of optical flow from different camera motions, a composite motion generates flow traces deformed from that of the simple camera motions. It is noticeable in Figure 4 that the flow mostly has a consistent direction (leftward) inverse to the camera motion direction, even if the image velocities (trace orientations) vary. According to step 2 of our algorithm, the designed slices for composite motions are planes or curved surfaces across the flows as indicated by dashed lines in red. The slice passes all the traces once to include the scenes stably appearing in the video clip.



S: static camera, T: translation, Z: zoom, R: rotation. $(T+R)_I$: around object motion with camera facing inward. $(T+R)_O$: the same motion with camera facing outward. S_T, S_D: directional and diversified motion of targets taken by a static camera, respectively

Figure 4. Possible cutting trajectories (red dotted lines) for major motion (traces in grey belts) in the condensed images. The downward vertical axis is the time axis. The camera motion is rightward if it is not static or zooming.

4. EXPERIMENTS

When a video is read in, the color is condensed to two images, in which we segment video to clips according to the smooth camera motions. In the condensed image, edges orthogonal to the time axis leave clear footprint of clips. Based on the gradient values, horizontal and vertical components of the major flow vectors are computed and compared. This information is then used in selecting orientation of the sampling slice and its scanning direction. The slice is generated from the video volume accordingly. The test video clips are from YouTube and other video sharing websites. Figure 5 shows several profiles from the composite motions. Many TV programs are the concatenation of clips from static cameras. If the clip is short, the resulting profile is not very different from the key frame as in (a). The profile has a higher temporal resolution than key frame because the diagonal slice is longer than the frame in width. (b) is the diversified motion in a static view. Traces of people are displayed in the condensed image. The profile shows walking area crowded by people. The major flow vector is small for the static camera so that both horizontal and vertical cutting can be considered. We select a vertical slice for a profile. The same selection is performed on zooming in clip (c). The part zoomed up is enlarged and one side of scenes is also listed in the profile. If the cutting curve is bended in the opposite way, it will create reoccurrence at the other side of scenes in the profile. (d) is the zoom during pan. (e) and (f) are profiles from camera pan. (g) is an around-object video taken by a camera on a moving car. Such a camera orbiting around a target is also frequently adopted in shooting a static object such as sculpture or exhibit on a table, performer on a stage, etc. The camera motion contains translation and rotation simultaneously. (h) is a rotating object in a static background, which is similar as an around-object video. Object points rotate with respect to the camera and show the flow as segments of ellipse in the video frames. Their traces in the condensed image are sinusoidal curves [15]. The resulting profile is sharp and complete with a squeezed background. The foreground is extended to show its aspect views that are impossible to be covered by a perspective image or key frame.

5. CONCLUSION

This work proposed a general framework to cut video volume to obtain a profile of video. It is based on the analysis of camera kinematics and motion flow in the condensed images. A uniformed algorithm has been implemented on videos from simple and combined camera motions, which theoretically guarantees the profiles from various video clips. The significance of the work lies in generating new video profile with both spatial and temporal information. Further, the method is for general camera motions and is more robust than conventional mosaicing and onion-skinning. It can facilitate video browsing and editing when automatically applied to a video database.

6. REFERENCES

[1] Janvier, B., Bruno, E., et. al. Information-theoretic temporal segmentation of video and applications: multiscale keyframes selection and shot boundaries detection. *Multimedia Tools and Applications*, 30, 3 2006), 273-288.

[2] Taniguchi, Y., Akutsu, A., Tonomura, Y. PanoramaExcerpts: extracting and packing panoramas for video browsing. *ACM Multimedia* 1997, 427-436.

[3] Bartoli, A., Dalal, N. and Horaud, R. Motion panoramas. J. *CAVW*, 15, 5 2004, 501-517.

[4] Liu, F., Hu, Y. and Gleicher, M. L. Discovering panoramas in web videos. *16th ACM Multimedia* 08, 329-338.

[5] Pritch, Y., Rav-Acha, A. Peleg, S. Nonchronological video synopsis and indexing. *IEEE PAMI*, 30(11), 2008.

[6] Correa, C. D., Ma, K. L. Dynamic video narratives*. ACM SIGGRAPH 2010*, 29, 4.

[7] Zheng, J. Y. Digital route panoramas. *IEEE Multimedia*, 10(3) 57-67, 2003.

[8] Zheng, J. Y., Tsuji, S. Generating dynamic projection images for scene representation and understanding. *Computer Vision and Image Understanding*, 72, 3 1998, 237-256.

[9] Zomet, A., Feldman, D., Peleg, S. and Weinshall, D. Mosaicing new views: The crossed-slits projection. *IEEE PAMI (*2003), 741-754.

[10] Flora, G. R., Zheng, J. Y., Adjusting route panoramas with condensed image slices. *ACM Multimedia, 2007.*

[11] Zheng J. Y., Bhupalam, Y. and Tanaka, H., Understanding vehicle motion via spatial integration of intensities. *ICPR*08.

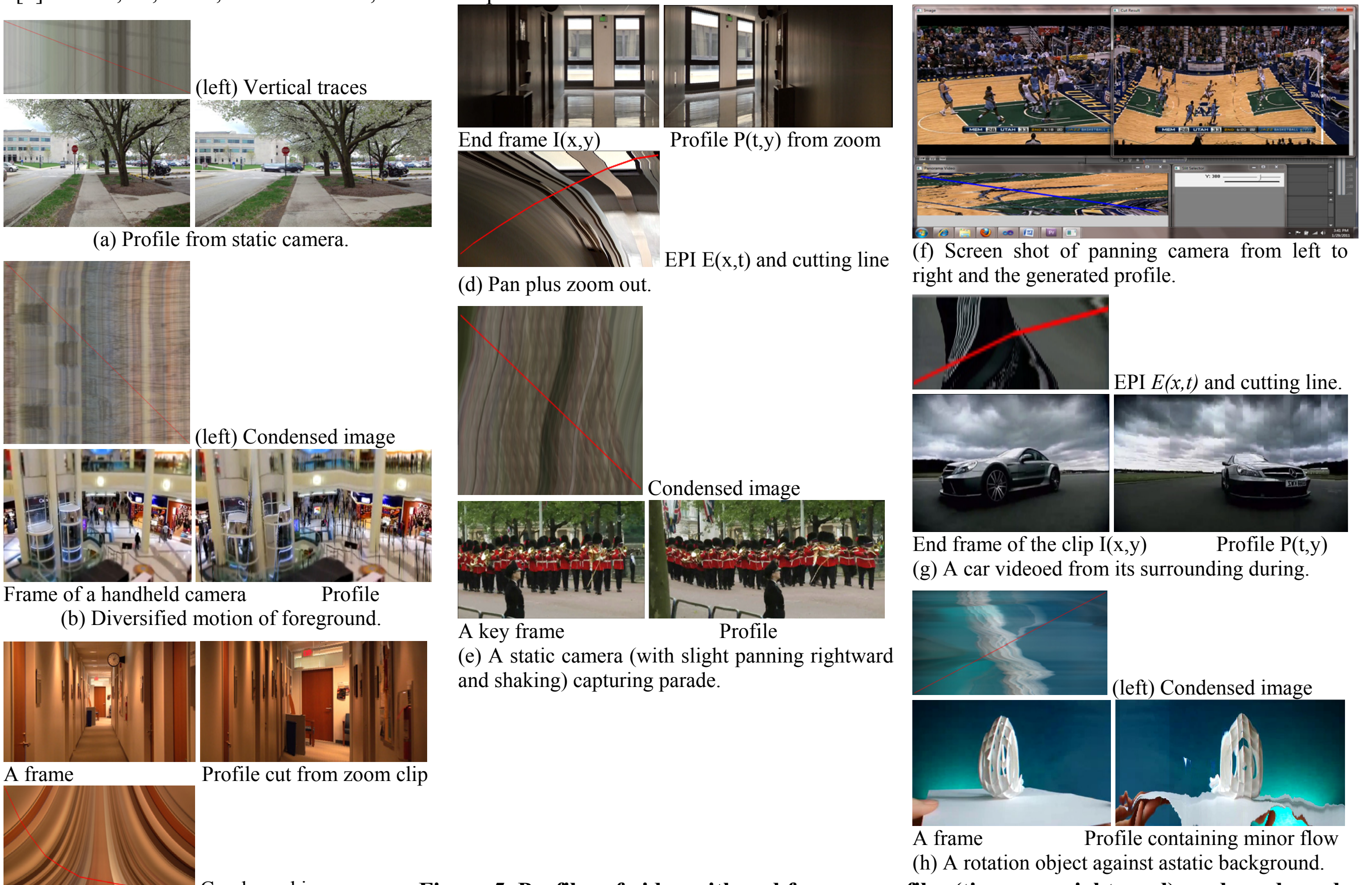
[12] Cai H., Zheng J. Y., Tanaka H. T., Acquiring shaking-free route panorama by stationary blurring. IEEE ICIP 2010: 921-924.

[13] Zheng J. Y., Cai H., Prabhakar K., Profiling video to visual track for preview, IEEE ICME 2011, 1-6.

[14] Zheng J.Y., Shi M., Scanning depth of route panorama based on stationary blur, IJCV, 78(2-3), 169-186, 2008

[15] Zheng, J. Y. Acquiring 3D models from sequences of contours. *IEEE PAMI,* 16(2),163-178, 1994.

[16] He, L., Shum, H. Y. Rendering with concentric mosaics. *SIGGRAPH*99, 299-306.


(left) Vertical traces

(a) Profile from static camera.

(left) Condensed image

Frame of a handheld camera Profile

(b) Diversified motion of foreground.

A frame Profile cut from zoom clip

Condensed image

(c) Profile from zoom clip.

End frame I(x,y) Profile P(t,y) from zoom

EPI E(x,t) and cutting line

(d) Pan plus zoom out.

Condensed image

A key frame Profile

(e) A static camera (with slight panning rightward and shaking) capturing parade.

(f) Screen shot of panning camera from left to right and the generated profile.

EPI *E(x,t)* and cutting line.

End frame of the clip I(x,y) Profile P(t,y)

(g) A car videoed from its surrounding during.

(left) Condensed image

A frame Profile containing minor flow

(h) A rotation object against astatic background.

Figure 5. Profiles of video with end frames, profiles (time axes rightward), and condensed images (time axes downward).