FINDING AND PRESENTING VIDEO DUPLICATES

S H Srinivasan

Applied Research Group
Satyam Computer Services Ltd.
SH_Srinivasan@satyam.com

ABSTRACT

With the proliferation of hand-held video camcorders, it is highly likely that several viewpoints of same scene are present in a home album. A mechanism is needed to detect, align, present, and summarize the multiple versions of the same scene or event. In this paper, we outline a framework for detection and alignment of video clips using ideas from image matching and multiview vision. We also discuss a technique for creating presentations and summaries out of aligned clips using attention models.

1. INTRODUCTION

Video clips are becoming common place with the widespread use of handheld video camcorders, webcams, mobile phones with video cameras, etc. With video cameras being available on personal devices, it is likely that the same event or scene is captured by several cameras at the same time. The different clips can have overlapping content. Because of sharing between users, several versions of the same event may be present in the *same* home album. Usually the file names and timestamps are the only clues to the content of the clips since metadata efforts like MPEG-7 are not widely deployed yet. File names and timestamps are often unreliable and misleading. Hence, without automatic organization, viewing multiple clips which may or may not be related, can be either frustrating or boring. Hence it is desirable that the related clips be automatically detected. We call several viewpoints of the same scene as *duplicates*. Duplicate detection ensures that different versions of the same scene/event can be clustered together. When presenting to the user, several versions of the same event should be presented. But making a consolidated presentation is more desirable. See figure 1 for a schematic of the application scenario.

Creating a consolidated presentation requires frame-to-frame alignment. Since different clips can be captured at different camera angles and motion, matching requires the application of techniques from multiview computer vision. Multiview computer vision techniques are computer intensive and not very robust when used in unconstrained environments. In this paper, we provide framework for creating presentations out of multiview clips using ideas from content-based image retrieval, multiview vision, and multiview attention theory. In particular, we perform the following steps.

1. Coarse-grained (shot-to-shot) matching of the clips.
2. Fine-grained (frame-to-frame) alignment of the matched shots.
3. Selection of the most appealing clip if several clips are available for a given event/scene.
4. Summarization of multiview videos.

This paper is organized as follows. Section 2 describes related work. Detection, alignment, and presentation are discussed in sections 3, 4 and 5. The paper closes with a discussion.

2. RELATED WORK

To the best of our knowledge the problem of organizing video clips into presentations has not been addressed in the literature. The subproblems have been addressed to different degrees of generality.

[1] discusses the detection of copies of video. The copies are “exact” in the sense that there is no change of camera angle or temporal origin. The distortions are due to encoding schemes and picture sizes. The work compares various features and distance measures. It is concluded that local edge representation gives the best results. In several application scenarios, the video are shot at arbitrary camera angles. Hence matching should be more flexible.

Three dimensional computer vision [2] has developed several models for camera imaging and matching procedures. The most general model – perspective projection – is also difficult to work with. Affine projection model is both realistic and tractable. [3] defines affine features and develops a matching technique for images. This is later used to match keyframes of shots of a movie [4]. Non-identical duplicate detection is investigated in [5].

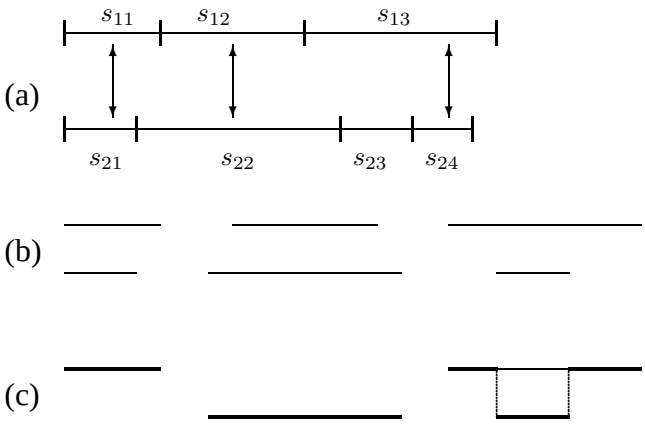

Fig. 1. Alignment and presentation of video clips. This figure considers only two clips c_1 and c_2. (a) Shot boundaries are identified. In this case, clip c_1 consists of shots s_{11}, s_{12}, and s_{13}. Clip c_2 consists of shots s_{21}, s_{22}, s_{23}, and s_{24} The key frames are matched setting up initial correspondence. In this example, the matched shots are (s_{11}, s_{21}), (s_{12}, s_{22}), and (s_{13}, s_{24}). Here clips are not organized on a common timeline. (b) The gross correspondence is refined based on motion information resulting in frame-to-frame alignment. Here shots are organized on a *common timeline*. (c) A presentation is created by choosing shots which are perceptually important. In case the shorter shot has more perceptual value, it is integrated into the longer shot using appropriate transition effects. In this case, s_{24} is integrated into s_{13}. Other presentation effects can be used.

All the above techniques are applicable to static scenes. Video contains motion - both object and camera. The system should identify object motion and should be insensitive to camera motion. [6] studies the spatio-temporal alignment of sequences. The cameras are either stationary or jointly moving. This constraint is not satisfied in several cases. [7] proposes a technique which works even with arbitrary camera motion.

The problem of attention has only been addressed recently in multimedia. [8] proposes an attention model using color, motion, and audio features. This model is now used across the video clips.

3. DETECTION

Detection of duplicates is by far the most computationally complex part of the system. Since video clips can have a large number of frames, a frame-by-frame comparison is impossible. Hence we perform three-level analysis. In the first-level, shot boundaries are identified and key frames of shots are extracted. In the second-level, the key frames are matched using a color-based criteria. A more exhaustive

Fig. 2. Key frames of two video sequences recorded at NUS campus. The first row shows key frames of first sequence which contains several scenes – including a desk (second in the first row). The second row shows the key frames of the second sequence which contains pictures of the desk at various angles. The matching should be invariant to camera angle.

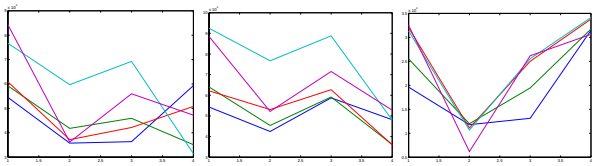

Fig. 3. Results of histogram matching. The histograms used are in RGB, HSV, and CIE XYZ space respectively. Since we are interested only in color matching, only HS and XY values are used. The histogram sizes are 5×5 (RGB) and 8×8 (HS and XY). The X-axis contains the key frame numbers of the first sequence. The matching scores are plotted for five key frames of the second sequence (four of which are shown in figure 2). All the key frame of the second sequence have similarity with only the second key frame of the first sequence. Hence the matching scores should be low only for the second frame. This is true only for the CIE XY space (third plot).

matching is performed on candidate color matches.

Shot detection and key frame extraction is based on [9]. Once the key frames of various shots are identified, they need to be matched. The keyframes are matched based on color features. We have explored several color spaces and found that CIE XYZ color space provides best results. Figure 2 show the keyframes of a video of a static scene we recorded ourselves. The clips show multiple views of the same scene. The matching scores for RGB, HSV, and XYZ color representations are shown in figure 3. XYZ space provides the best result.

The candidates found by the color matching algorithm are only approximate and need to be refined. This is done using by establishing wide baseline correspondence between the candidate matches. We use the technique proposed in [10].

4. ALIGNMENT

Frame-by-frame alignment is impossible if there is no motion in the scene. Several algorithms have been proposed for spatio-temporal alignment of sequences. In this paper we use the technique proposed in [7]. This technique tracks a collection of points (corners) forming trajectories in the two clips. Let the corresponding points in the two trajectories be (x_i, y_i) and (u_i, v_i). If there is a perfect match between these points, the following equation is satisfied.

$$\begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}^T F \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = 0$$

where F is the 3×3 Fundamental matrix. This can be rearranged in the form

$$f^T d = 0$$

where

$$d = [x_i u_i \ x_i v_i \ x_i \ y_i u_i \ y_i v_i \ y_i \ u_i \ v_i \ 1]^T$$

and

$$f = [f_{11} \ f_{12} \ f_{13} \ f_{21} \ f_{22} \ f_{23} \ f_{32} \ f_{32} \ f_{33}]^T$$

The vector f consists of f_{ij}s and the data vector contains the data terms. For several data points, the second term can be organized into a matrix and the equation becomes a matrix equation as shown below.

$$f^T D = 0$$

If there are N data points, the size of D is $9 \times N$. Since the equation is homogeneous, the rank of the matrix is 8. Hence the size of the ninth singular value is a measure of the error in the data. If the ninth singular value of the matrix is large, a mismatch is signaled. Hence the ninth singular value can be considered to used as a distance metric between trajectories.

The above alignment uses only visual information. Audio information can be used to check the above alignment. We assume that frame-based alignment of video also results in corresponding fine-grained alignment of audio. We calculate the correlation coefficient of the instantaneous sound energies of the two clips. If the correlation coefficient is less than a threshold, then the match is rejected.

5. PRESENTATION

Instead of presenting a collection of shots to the user, a consolidated presentation is temporally compact and aesthetically more pleasing. Some users may prefer maximum length presentation while other may like to view a summary. Summarization issues are discussed in the next section. To create a presentation, we compute the user attention score of each frame of video according to [8]. The shot matching enables us to arrange the clips on a common time line. For timelines for which multiple shots are available, we calculate the average and peak difference of attention values. If the difference of the average is high, the clip with higher attention value is chosen. If the averages are almost equal, then the frames from the shots are chosen depending on their attention score with a minimum duration constraint.

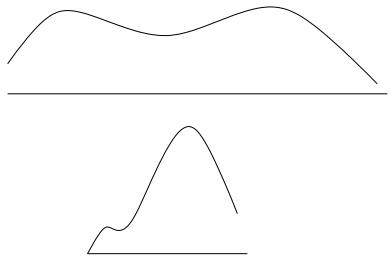

Fig. 4. Multiview attention. The figure shows the visual attention values of two temporally overlapping clips. It can be seen that the second clip has higher attention value though it is temporally shorter. Hence the final presentation is obtained by splicing the two clips. See also figure 1.

Let s_1 and s_2 be two temporally overlapping shots with s_2 being the shorter one. We calculate the visual and audio attention profiles of the two shots. Video and audio are handled separately for the following reasons.

1. For a given shot, the audio may have its own sentence boundaries – indicated by pauses. Hence a fine-grained mixing of audio is possible.
2. For video, we do perform mixing of shots - under certain conditions. It is possible to use certain transition effects for this. For audio, no transitions are used.

Let us first consider the video attention model. Figure 4 shows the attention profiles of two shots. When we consolidate the shots, we always use the longer shot. The question is whether to use the short one. Let $\bar{v}_1$ and $\bar{v}_2$ be the *average* visual attention values and v_1 and v_2 be the maximum attention values for the overlapping part. Then s_2 is used for the overlapping part if

$$\bar{v}_2 - \bar{v}_1 > \tau_1 \text{ and } v_2 - v_1 > \tau_2$$

where τ_1 and τ_2 are thresholds. By using both average and maximum values, the chances of induced shot boundaries is minimized. When a shot is spliced, the transition is usually abrupt. The camera angle also changes abruptly. It is possible to smoothly change the camera angle. This requires the *synthesis* of intermediate views. Since at least two views are available, view synthesis is possible in principle. For

some scenes and object configurations, synthesis may not be possible. View synthesis leads to aesthetically pleasing results.

The audio for a shot is segmented into sentence boundaries based on energy condition. Because of the correlation test performed, we can be sure that the sentence boundaries are synchronized in the two clips. We choose the sentence which has maximum attention value.

5.1. Summarization

Home videos tend to be long. If several clips are integrated together, the presentation is likely to be longer. The previous section produces a single video clip which contains the relevant information. We can apply summarization techniques [8] on this single video. Since visual and auditory attention values are already available, summarization does not require additional computation.

6. DISCUSSION

In this paper we have addressed the important problem of detecting, aligning, presenting, and summarizing multiple video clips of the same event. The solution involves ideas from image matching, motion matching, and attention. A prototype based on the ideas discussed is being built.

7. REFERENCES

[1] A Hampapur and R Bolle, "Comparison of distance measures for video copy detection," in *IEEE ICME*, 2001.

[2] O Faugeras and Q-T Luong, *The Geometry of Multiple Images*, MIT Press, 2001.

[3] F Schaffalitzky and A Zisserman, "Multi-view matching for unordered image sets, or "how do I organize my holiday snaps?"," in *European Conference on Computer Vision*, 2002.

[4] F Schaffalitzky and A Zisserman, "Automated scene matching in movies," in *Proceedings of the Challenge of Image and Video Retrieval*, 2002.

[5] A Jaimes, S F Chang, and A C Loui, "Duplicate detection in consumer photography and news video," in *ACM Multimedia*, 2002.

[6] Y Caspi and M Irani, "Spatio-temporal alignment of sequences," *IEEE PAMI*, vol. 24, no. 11, pp. 1409–1424, 2002.

[7] C Rao, A Gritai, M Shah, and T Syeda-Mahmood, "View-invariant alignment and matching of video sequences," in *International Conference on Computer Vision*, 2003.

[8] Y F Ma, L Lu, H J Zhang, and M Li, "A user attention model for video summarization," in *ACM Multimedia*, 2002, pp. 533–542.

[9] H J Zhang, A Kankanhalli, and S W Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, 1993.

[10] S H Srinivasan and M Kankanhalli, "Wide baseline multi-frame matching using random walks," in *Asian Conference on Computer Vision*, 2004.