ICONIC VERSUS NATURALISTIC MOTION CUES IN AUTOMATED REVERSE STORYBOARDING

R. D. Dony*, J. W. Mateer, J. A. Robinson, M. G. Day

*University of Guelph, School of Engineering, Guelph, Ontario, Canada N1G 2W1,
University of York, Department of Electronics, Heslington York YO10 5DD UK
Contact: John Robinson, jar11@ohm.york.ac.uk

Keywords: Video analysis, summarisation, visualisation.

ABSTRACT

Storyboarding is a standard method for visual summaristion of shots in film and video preproduction. Reverse storyboarding is the generation of similar visualizations from existing footage. We have previously demonstrated that the use of storyboarding techniques, such as *trail lines*, *onion skins* and *streaks* in conjunction with *mosaic constructions*, can succinctly summarise many visual sequences. However, there are instances involving complex subject and/or camera movement where such approaches are not successful. We identify specific types of these that require alternative summarisation strategies and propose new methods using graphic arrows to accommodate these cases effectively. The generation and rendering of arrows indicating subject motion within frame, and the determination and application of arrows to indicate intermediate camera motions are discussed in detail. These techniques result in clear and succinct representations of complex camera and subject motion by breaking sequences into easily interpretable parts.

1 Introduction

The summarisation of film, video and television content is vital to many production, archival and analysis tasks. Workers in these areas require access to specific details or attributes of footage without having to refer to the source material. This information should be presented in a consistent form utilising standard terms and representations commonly found within the domain.

In media production industries, *storyboards* – static two-dimensional graphic depictions of content of each shot and scene – are widely used for visual summarisation. Tasks such as set design, location lighting and image compositing are made more efficient through the availability of a common reference to the 'vision' of the piece. Shorthand descriptions of all important visual components of each shot provide clear and accurate depictions of motion sequences in static form. Many researchers working in the area of automated media analysis use the term storyboard to denote a series of still images or keyframes extracted from a visual sequence to describe its content. Production storyboards – those used by content producers – are much more detailed. They are used in numerous preproduction, production and postproduction contexts when a succinct visual summary of what is to appear on screen is required. In this paper we use the term in its production context.

To provide the detail required by practitioners for production, storyboards incorporate the following types of information:

1. Composition of the overall shot
2. Appearance of foreground elements
3. Indication of object movement
4. Indication of camera movement

The first two types are typically conveyed through simple drawings rendered in the correct aspect ratio of the medium used (e.g., 16:9 for widescreen television, etc.) These drawings can be either black and white or full colour. The amount of detail included will be dictated by the importance of elements to the production of the shot. For example, if the purpose of a shot is to show that a character has a gun in their pocket, the drawing should show a suitable bulging shape in the fabric to make this clear to the viewer (and production team). Insignificant elements can be shown in a very rough form or excluded entirely.

Hollywood has developed specific drawing techniques to convey the last two types of information. *Onion skins* (multiple instances of a subject that indicate specific intermediate positions), *streaks* (lines that show the trajectory of a moving object) and *trail lines* (repetitions of the trailing edge of a moving object to indicate the speed of motion of the moving object) are long-established summarisation methods, similar to those used in comic books. *Field cuts* (outlines of intermediate camera positions in the correct aspect ratio) and *mosaics* (panoramic views showing the area caught on camera) are used to denote important aspects of composition and content during camera movement.

In our previous work [1] we argued that the storyboard metaphor, which is highly effective for preproduction visualisation, could be equally effective as a means of summarising visual media content. We developed *Automated Reverse Storyboarding*, a set of computational methods to create visual summaries of existing video footage automatically using *mosaics*, *onion skins* and *trail lines* in keeping with the production

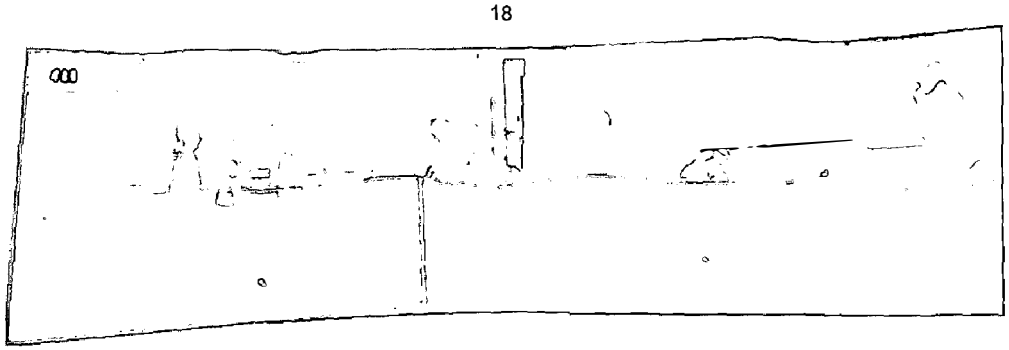

Figure 1: Naturalistic reverse storyboard illustrating mosaicing, onion skins, trail lines and field cuts

storyboard metaphor. In many circumstances, where there is object and/or camera motion in a single direction, these prove very effective at concisely describing shots. Figure 1 shows an example.

However, in highly complex shots, where subject motion overlaps or camera motion occurs along multiple axes, results from these techniques can be unsatisfactory. (Some examples are shown in later figures.) This leads us to consider using motion arrows, which are the last important visualization cue of preproduction storyboarding. Figures 2 and 3 show examples of their use for both object and camera motion visualization.

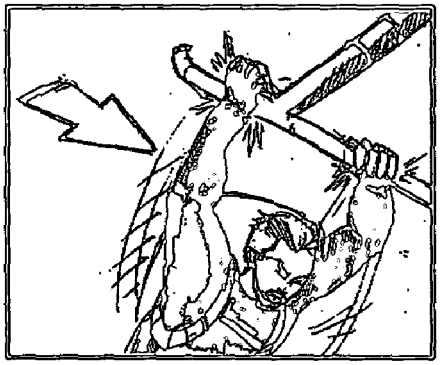

Figure 2: Arrows indicating subject motion

Arrows and field cuts are "iconic" ways of conveying object motion. They do not augment the shot imagery in a naturalistic way, by merging semi-realistic visual cues into the scene, as streaks and trail lines do. Rather they overlay the visual material, clearly distinct from it. Although we have used field cuts previously, our storyboards have otherwise been naturalistic. We therefore consider in this paper the generation of boards at the opposite end of a naturalistic/iconic continuum: using arrows for both camera and object motion cues, dispensing with the integrative shot mosaic as well as with trail lines and onions skins. Later we consider how the two approaches may be merged into storyboards that combine naturalistic and iconic cues, just as preproduction boards do.

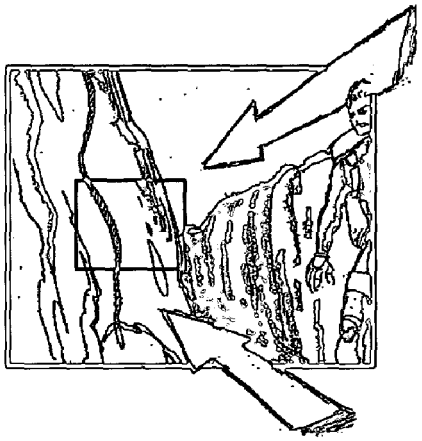

Figure 3: Arrows indicating camera motion

2 Prior Work

Katz provides an excellent description of how storyboards are created and used by practitioners [2]. However, this process has not been widely considered for automated summarisation outside of Mackay and Pagani's work [9] and our own [1,3]. The generation and overlaying of cartoon-style motion cues, widely used by industry storyboard artists, has been examined [10] but relies on the user identifying specific objects or areas of interest within frames. As of this writing we are unaware of any work examining the role of arrows in describing either subject or camera movement within shots.

As discussed below, our new technique overlays arrows on one or two keyframes. The extraction of keyframes as a means of content summarisation has been widely studied. Many researchers have focused on establishing which single shot best conveys a sequence (such as [4, 12-14]) whilst others have centred on finding more intuitive ways to present those frames (including the determination and subsequent larger display of dominant frames [5] and the use of the Japanese comic book-inspired Manga layouts described in [6]). Other work has looked at different forms of video abstraction (including [7]) and appreciable effort has been expended to enable accurate automated summarisation (an example is [8]).

We now briefly review our prior work on reverse storyboarding. The initial stages – parsing into shots and preliminary analysis of movement within the shot – are

common to the earlier work and the new system reported here.

In order to effectively summarise film or video content using storyboarding metaphors we must first categorise shots based on other parameters. Using *ASAP*, an automated shot analysis program [11], we segment visual sequences into shots and then classify the camera movement contained within each shot (i.e., pan, tilt or zoom). ASAP also provides frame-to-frame projective transform estimates. By composing these, we can generate shot mosaics, which, in our earlier work, were used as the background for all storyboards.

If a shot is static (i.e., a 'hold') and there is no motion or simple subject motion, such as movement in one direction or a small repetition within a confined area of the frame then the use of onion skins and trail lines can be very effective. Figure 4 shows an example from the earlier system. Note how both start and end framings are marked even though this shot is almost a hold.

Figure 4: Onion skins showing repetitive motion

If camera motion is present and does not change direction along a single axis, then mosaics with field cuts present a clear depiction of the shot as shown in figure 5.

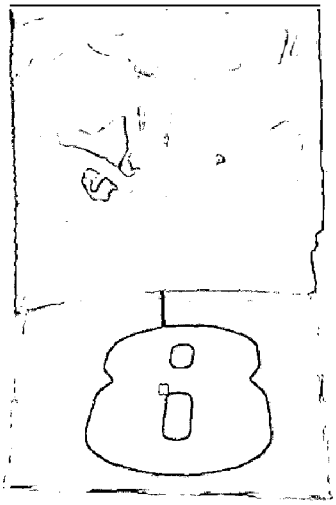

Figure 5: Mosaic description of tilt up

Note, the camera work can be fairly complex including multiple moves, such as a pan right followed by a tilt up with a zoom in, so long as the motion does not double back upon itself. Such an example is shown in figure 6.

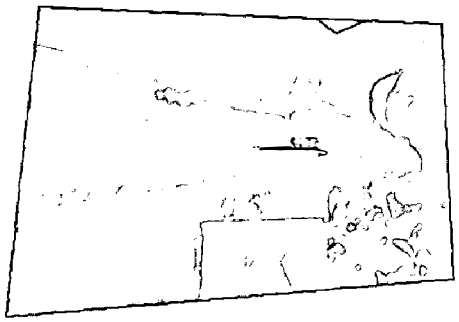

Figure 6: Mosaic description of compound camera move

If camera motion and subject motion are both present, onion skin, trail line, field cut and mosaicing techniques can all be used with good effect as illustrated in figure 1.

When camera motion is complex, the shot is segmented into subshots at the boundaries of major camera motion. Some examples of this are shown in later figures.

The examples above all illustrate that naturalistic visualization works for many simple shots. However, examples shown later illustrate its limitations.

3. An arrow-based reverse storyboarding method

We propose an alternative to naturalistic reverse storyboarding that relies on just one or two keyframes overlaid with graphical elements (arrows) to convey all the motion information in the shot. As in our earlier system, we begin by parsing camera motion, segmenting a shot into subshots if necessary. We then determine whether a subshot involves a substantial camera move. If so, we use two keyframes; if not we use only a single frame. Arrows are then derived from the camera motion estimator and a foreground-object analyser as described below.

Keyframe selection

In common with previous researchers we have investigated keyframe selection based on spatial and temporal features of the frames within a shot. We developed a statistical classification strategy that used combinations of motion vectors and image histograms. We also implemented a prior-art keyframe selection algorithm [13], plus a middle-frame selector.

In evaluating keyframe extractors we compared their outputs with preproduction boards as one method of assessing performance. In this process we noted that the distribution of keyframes as selected by *humans* within a shot is approximately uniform. It appears not to have been argued previously that if closeness in time correlates with image similarity, then a frame taken from the middle of the shot has highest expected similarity with the true keyframe. In our experiments, we found no

significant advantage for a complicated keyframe selection algorithm over choosing the middle frame, and we therefore adopt the middle frame when the move is a hold or contains minor camera movement.

When a shot or subshot contains a substantial camera move we must represent the extent of that move accurately. Preproduction storyboards always show start and end framing when the two are radically different, and we therefore use the first and last frame to represent the range of the shot. These first and last keyframes are arranged relative to the direction of camera motion. For example, in the first part of figure 11, the camera pans right so the last frame is positioned to the right of the first. With the addition of an arrow indicating the direction of camera motion, the result can succinctly convey the essential camera motion of many shots.

Foreground Elements – Arrow generation

Once keyframes are identified, motion cues of objects within the shot are then drawn using arrows. One approach to identify moving foreground objects in a shot is to employ object tracking techniques [15]. This topic in computer vision has been extensively investigated [16-19] especially in the context of human motion. The goal of such techniques typically is for measurement and modelling. Further, many require user assistance and can operate in only constrained or simplified environments. The goal of our work is to produce a visual representation of the shot in a fully-automated and general-purpose manner. Therefore such techniques are not suitable for our purpose.

We instead turn to the method previously developed in [1]. In this approach, interframe projective transforms provide an initial spatial registration for all frames in the shot, then the time series of luminance values for each registered point are collected into histograms. Time medians derived from the histograms are taken to correspond to the luminances of the static background. The approach is intuitively reasonable since the pixel intensities of moving foreground objects are less likely to occur than the intensity of the static background, and it proves successful in practice for a large proportion of shots with moving objects. Having determined background values, the foreground objects are simply identified as those regions in a frame that differ significantly from the background.

Figure 7 shows the results of such a segmentation for 2 frames from a hold shot of a camera crew as shown in figure 8. Note that the microphone boom and moving person in the lower right are well segmented from the background. As the crew members do not move much during the shot, they are considered as background by the median operator. They therefore do not show up under this segmentation scheme. These two frames are from the middle portion of the shot and are four frames apart. The motion of the objects, namely the microphone boom and the moving person, are quite apparent.

The moving objects are thus identified for the keyframes.

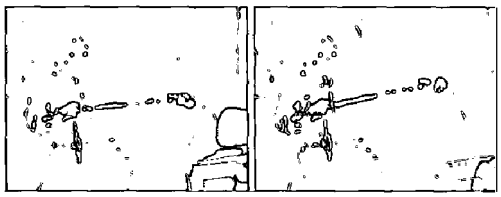

Figure 7: Segmentation of two frames, four frames apart from the middle of a hold shot. Note the segmentation of the microphone boom in the centre of the frames and the moving person in the lower right corner.

To determine the motion of the objects, object identification is performed on a number of frames previous to the keyframe. The segmentation maps are processed via morphological operators to remove spurious regions and fill in gaps. The individual objects are identified and tracked through each of the n preceding frames. Segmented regions are considered as contiguous objects if some overlap exists between frames at any portion of the sequence. Throughout this tracking, the camera movement is accounted for through the use of the projective transform previously calculated. The tracking of the objects allows for splits and merges due to possible inaccuracies in the segmentation. Further, thresholding based on object size and movement is applied to filter out small objects. The first location of each object, j, is taken as its centroid of the first of the n frames preceding the keyframe and is calculated as

$$x_j = \frac{1}{N_j}\sum_{x \in O_j} x_i \qquad y_j = \frac{1}{N_j}\sum_{x \in O_j} y_i$$

Note that because of the need to establish the motion before a keyframe, in the case of a pan, the first keyframe is actually a number of frames after the first frame. Similarly, the last location of an object is calculated as the centroid of the object as it appears in the keyframe.

The first and last locations are used to draw the arrow representing the motion. The head of the arrow is spaced back from the object by its approximate radius calculated as

$$r = \sqrt{N/\pi}$$

where N is the number of pixels in the object. The length of the arrow is proportional to the Euclidean distance between the start and end points. For visual effect, a scaling factor of two was used, i.e. the length of the arrow is twice the amount of object movement. The width of the arrow is an indication of the size of the object. It is set to half the object radius as calculated above.

4 Results and Discussion

We present paired comparisons of shots summarised with keyframes and arrows as described in section 3, and with mosaics, trail lines and onion skins (our previous method). The figure captions provide some indicators of the tradeoffs, but the reader will observe that interpretation depends on shot content in a complicated way.

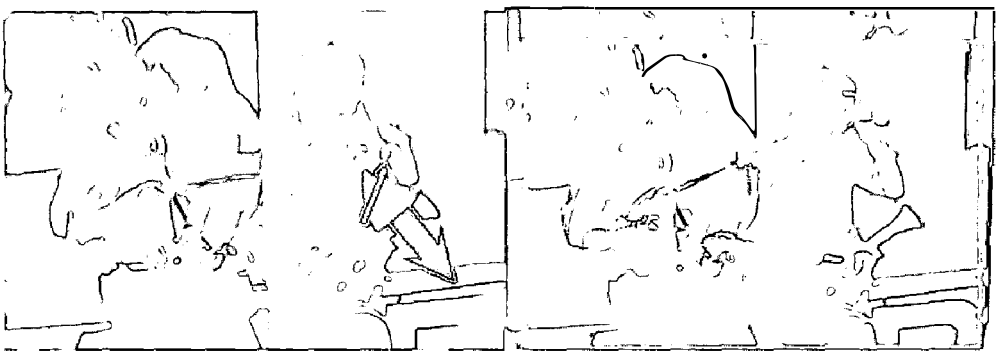

Figure 8. Two objects in motion within a hold. The arrows version gives a clearer representation of the moves. (This is the shot illustrated in figure 7)

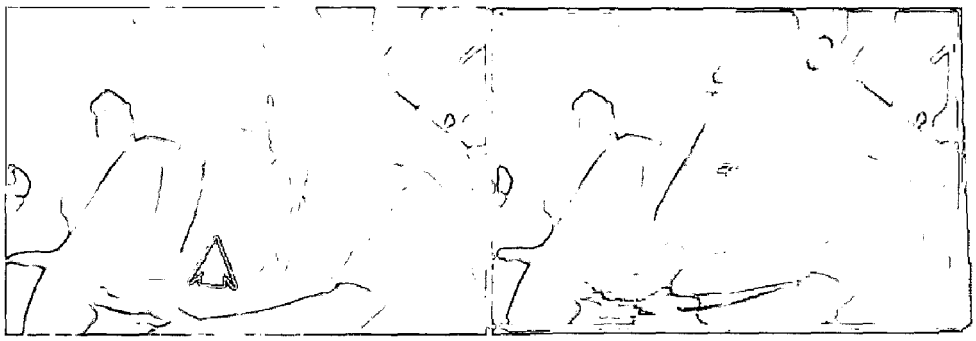

Figure 9. A hold with a single large movement (raising the arm) then a small repeated movement (brushing the hair). While the arrow provides the superior cue for the first, the trail lines suggest the second.

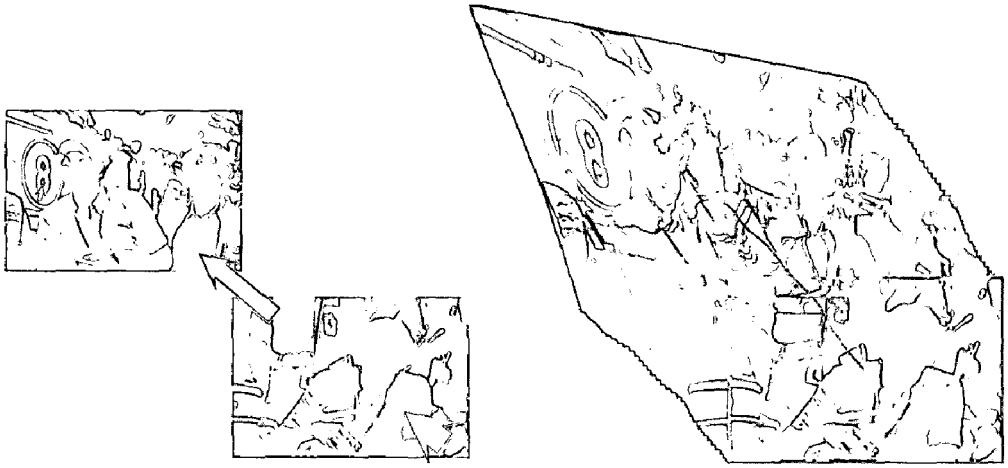

Figure 10. A pan/tilt with large object motion. In this case the mosaic generation is successful despite the large moving object in the frame. However the onion skin visualization is ambiguous and there are too many trail lines. The arrow version is again cleaner though the connection between the start and end frames would be better represented by aligning them more closely.

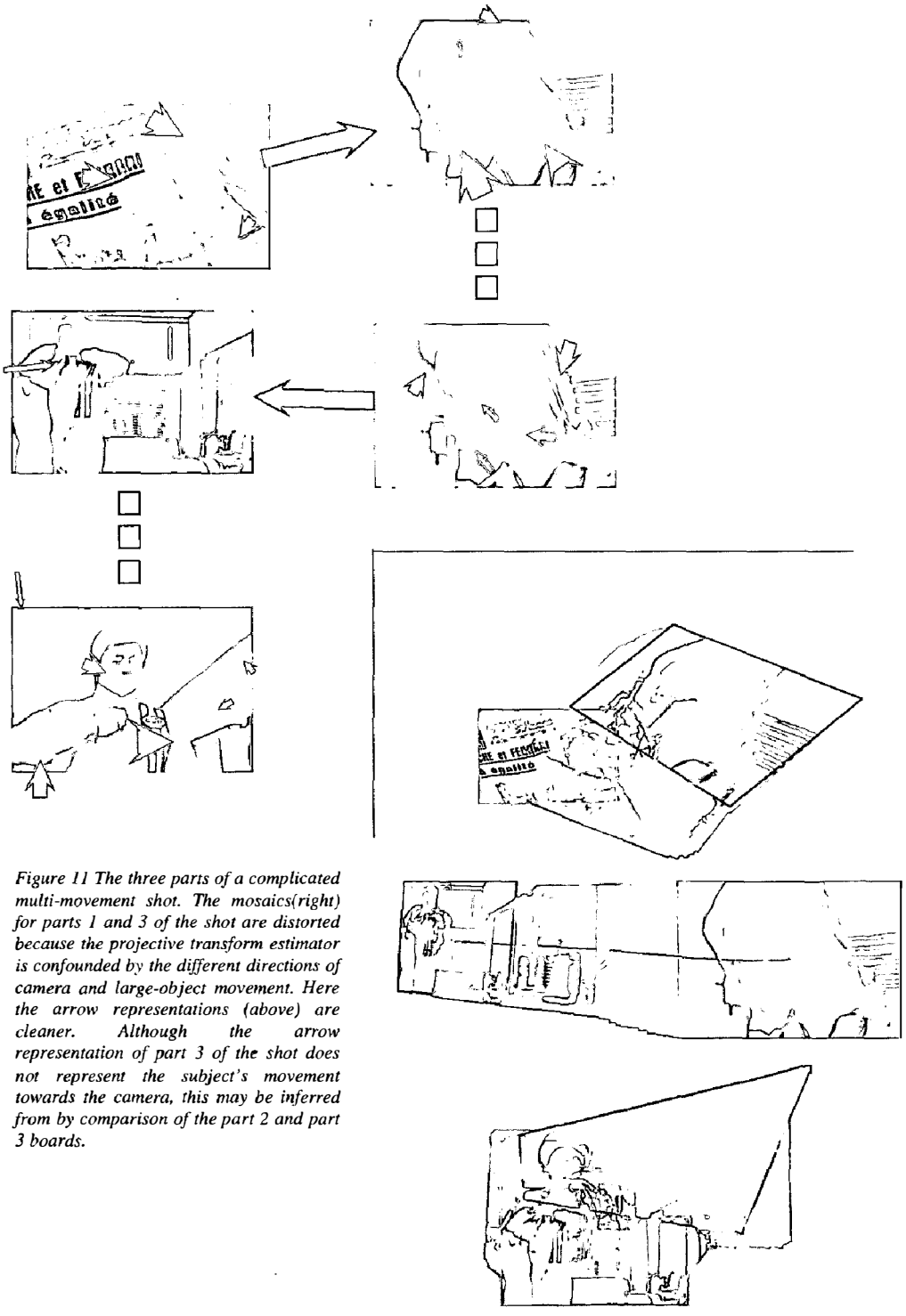

Figure 11 The three parts of a complicated multi-movement shot. The mosaics(right) for parts 1 and 3 of the shot are distorted because the projective transform estimator is confounded by the different directions of camera and large-object movement. Here the arrow representations (above) are cleaner. Although the arrow representation of part 3 of the shot does not represent the subject's movement towards the camera, this may be inferred from by comparison of the part 2 and part 3 boards.

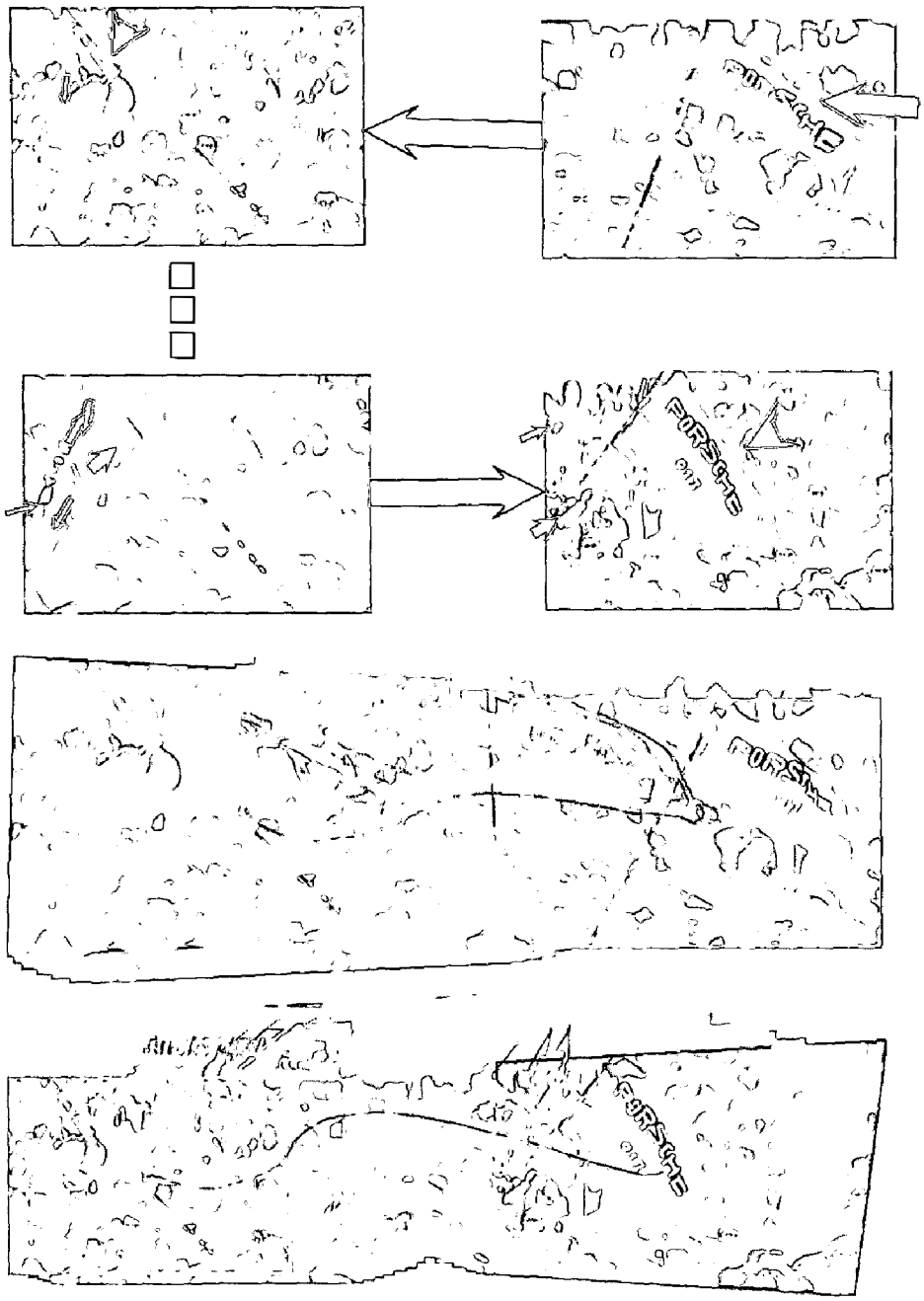

Figure 12. Another two-part shot with a waving flag. The subtleties of flag flap are captured better in the arrow version, but the mosaic provides better context over the whole shot.

As the figure captions discuss, there are tradeoffs between mosaics and arrows in representing both camera and object motion. Which is better depends on the shot content, suggesting either that an automated system for choosing between the two would be useful, or that combination of both would be better than either alone. The combination of both methods requires the modification of parameters so that storyboards do not become too cluttered. Results for two of the subshots shown previously are given in figure 13.

Figure 13. Two examples combining mosciaing, trail lines and arrows

The two examples above are wide pans and it is this type of shot for which the combination of visualizations is most effective. A storyboard shown earlier as figure 1 is made even clearer with addition of object motion arrows, as figure 14 shows.

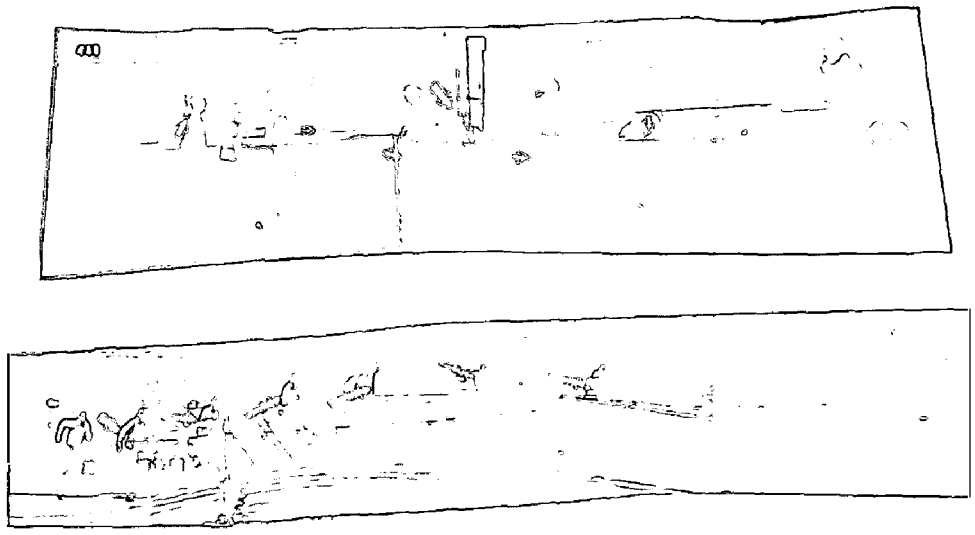

Figure 14. Pan mosaics enhanced with arrows

Based on the above results, we propose the following algorithm for generating storyboards that combine the strengths of naturalistic visualization and iconic representation via arrows.

1. Generate a shot or subshot mosaic by composing estimated frame-to-frame projective transforms.

2. if the positions of the first and last frames when warped into the mosaic do not overlap:

- use the mosaic for visualization, add trail lines, onion skins and field cuts as in [1] and add object arrows (yellow) as this paper.

else if the position of the last frame centre is outside the boundary of the first frame in the mosaic:

- use first and last frames as keyframes, position according to the dominant camera move and add camera (blue) and object (yellow) arrows as in this paper.

else:

- use the middle frame as a keyframe, and add object arrows (yellow) as in this paper.

This algorithm generates good composite storyboards for the 26 shots on which we have applied it. In future work we will generate composite storyboards for entire movies and identify the limitations of the algorithm for selecting and combining storyboard components.

5 Conclusions

We have developed an iconic reverse storyboarding system to complement our earlier naturalistic system. We have presented results comparing the two approaches, illustrating how each has superior properties depending on shot content. We have also showed how naturalistic elements (background mosaic, onion skins, trail lines) and iconic elements (object motion arrows, field cuts) can be combined in an integrated presentation. The results show that due to the rich variety of shot composition found in production footage, no single approach works for all cases. We have proposed an algorithm that chooses between the different visualisation methods investigated here.

References

1. Dony, R.D., Mateer, J.W., Robinson, J.A. (2005) Techniques for Automated Reverse Storyboarding, IEE Proc. – Vis. Image Signal Process., vol 152(4), pp. 425-436
2. Katz S D, (1991), "Film Directing Shot by Shot", Michael Wiese Prods/Focal Press, Stoneham, USA
3. Mateer, J. W. and Robinson, J. A (2003) "Semi-Automated Logging for Professional Media Applications", proceedings of Video, Vision and Graphics 2003, Bath UK, pp.25-31
4. Hammond, R & Mohr, R (2000) "A probabilistic framework of selecting effective key frames for video browsing and indexing", International workshop on Real-Time Image Sequence Analysis
5. Yeung, M.M, Yeo, B-L (1997) "Video visualization of compact presentation and fast browsing of pictorial content", IEEE Trans Cct & Sys for Video Tech, Vol. 7(5), pp. 771-785
6. Girgensohn, A (2003) "A fast layout algorithm for visual video summaries", proceedings of IEEE International Conference on Multimedia and Expo 2003, Baltimore, vol. 2, pp.77-80
7. Lienhart, R, Pfeiffer, S, & Effelsberg W (1997) "Video Abstracting", Communications of the. ACM, vol 40(12):, pp. 55-62
8. He, L., Sanocki, E., Gupta, A., & Grudin, J., (1999) "Auto-summarization of audio-video presentations", proceedings of ACM Multimedia'99
9. Mackay, W.E. & Pagani, D. (1994) "Video Mosaic: Laying out time in a physical space", Proceedings of Multimedia '94 . San Francisco, CA: ACM.
10. Collomosse, J.P., Rowntree, D., & Hall, P.M. (2003) "Cartoon-style rendering of motion from video", proceedings of Video, Vision and Graphics 2003, Bath UK, pp.117-124
11. Mateer, J. W. and Robinson, J. A. (2003) Robust Automated Footage Analysis for Professional Media Applications, proceedings of Visual Information Engineering 2003, Guildford UK, pp. 85-88
12. Dufaux, F. (2000) Key frame selection to represent video. Compaq Computer Corp., Cambridge Research Lab, ICIP00 (Vol II, pp. 275-278)
13. Liu, T. M., Xhang, H. J., Qi, F. H.. (2003) A Novel Video Key Frame Extraction Algorithm. Microsoft Research China, Beijing. Shanghai Jiaotong University, Shanghai.
14. Divakaran, A., Radhakrishnan, R., Peker, K. A., (2001) Video summarizarion using descriptors of motion activity: A motion activity based approach to key-frame extraction from video shots. Journal of Electronic Imaging vol 10(4), pp. 909-916
15. J. P. Collomosse, D. Rowntree, and P. M. Hall. (2003) Cartoon-style rendering of motion from video. In *Vision, Video and Graphics*, pp 117-124,
16. S. Blackman and R. Popoli. (1999) *Design and Analysis of Modern Tracking Systems*. Artech House
17. D. Bullock and J. Zelek. (2002) Real-time tracking for visual interface applications in cluttered and occluding scenarios. In *Proceedings of the 15th IEEE International Conference on Vision Interface (ICVI)*, p. 751
18. J. C. Clarke and A. Zisserman. (1996) Detection and tracking of independent motion. In *Image Vision and Computing*, vol 14, pp. 565-572
19. Richard D. Green and Ling Guan. (2003) Tracking human movement patterns using particle filtering. In *Proceedings IEEE ICASSP 2003*, pp. 25-28, Hong Kong.