CONTENT CONTROLLED IMAGE REPRESENTATION FOR SPORTS STREAMING

Anil Kokaram, Francois Pitie, Rozenn Dahyot, Niall Rea, Serge Yeterian

Signal Processing and Media Applications Group,
Electronic and Electrical Engineering Department,
College Green, Dublin 2, Ireland

ABSTRACT

Content based analysis has traditionally been posed in the context of identifying some material in response to a user query. This paper illustrates that given a content based analysis process that can identify semantic events in a sequence, that sequence can then be changed in various ways. A Motion Keyframe is presented to re-express the viewing of a sequence. The notion of content analysis for *control* of other media processing engines is introduced. Tennis footage is used to illustrate the ideas since sports in general contains strong contextual information.

1. INTRODUCTION

Content based analysis for visual media has traditionally been posed in the context of indexing/retrieval or summarization. The scenario is that in order to access information in non-text media, automated extraction of content based cues like colour, shape and motion, can yield more flexible access than is possible with manually input text metadata only. Recent work has emphasised the importance of the user context in bridging the semantic gap between feature extraction and manipulation and the interpretation of this information for the benefit of the user. This paper presents the notion that there are wider implications of this kind of technology for *content aware media processing* in general. It focuses on the sports genre because it is an example of high commercial and popular importance as well as providing a well established user context within which the semantic gap is bridgeable [6, 1, 9].

A first example, already explored in [3] points out that by detecting important events in a media stream it is possible to alter the bitrate allocated to the stream in proportion to importance. Thus at a high action, and semantically relevant event in sports, the bit rate allocated would allow the full frame rate motion to be viewed. For low activity events, still frames at a low frame rate could be transmitted. Thus in tennis for instance, each contact with the ball will result in high frame rate motion while otherwise a series of low frame rate stills would be transmitted. The idea is that this will allow high bandwidth sports events to be viewed over narrow bandwidth wireless links.

In fact that example is just one from many possibilities. In a broader context, consider the object based coding in MPEG4 or H.26x. Object segmentation from any video stream is a difficult and well researched task. It is well understood that given an arbitrary video stream it is very unlikely that objects of any semantic relevance could be reliably extracted from such a stream in a manner that would be useful for any kind of object based manipulation or editing. However, within a specified context a solution to the problem is possible. Consider sports footage of well defined games like tennis in which player position and camera orientation are well understood a-priori. This information can be used to determine the number and type of objects in each scene given that a content based view classification can be performed as in [6, 5]. In sports each view has specified content e.g. the court shot in tennis would contain the court and two players. This makes the segmentation problem tractable as information about the specific colour and motion of relevant objects can then be made available.

The idea of content aware media processing can be taken further. If events in sports can be identified and tagged with a level of importance, why not use this information to fundamentally change the nature of the image representation? The representation can be altered to match the user need. For instance, for viewing sports on low quality displays in poor lighting conditions (e.g. mobile devices), the semantically relevant objects can be highlighted or enlarged. In the case of presenting a clickable menu of scenes to navigate through an event, a keyframe image that represents motion as well as semantic objects in some intuitive way would improve the user interfacing. In media streaming, the images themselves could be re-expressed in a way that alters the bandwidth without affecting the semantics of the event. This is content adaptation in its broadest sense. In the next section, the notion of a Motion Based Key Frame is presented as an expression of content aware media processing. This kind of keyframe is appropriate in all the contexts

Work funded by Enterprise Ireland (MUSEDTV, CASMS), HEA and European Union (MOUMIR, MUSCLE).

ISBN: 952-15-1364-0

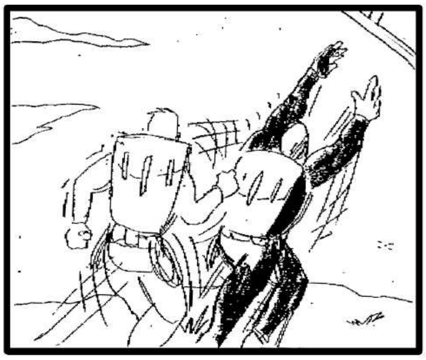


Fig. 1. Onion Skinning *to create the effect of motion in a still [7].*

discussed.

2. THE MOTION BASED KEY FRAME

Comic strip artists have long been used to conveying the notion of movement in a single image. Figure 1 is an example of a technique known as *onion skinning* and the effect is clear. That technique can be used to collapse semantic object behaviour over a series of frames, onto a single image. This kind of effect has been considered for gesture classification (the motion history image (MHI) [2]) as well as artistic purposes [10]. The paper **Salient Stills** [10] by Massey and Bender proposes to collapse frames into a single frame based representation. However, the authors rely on explicit segmentation and motion from MPEG decoding to create the effect. In practice the segmentation operation is never robust unless the context is known. In the context discussed in this paper, given the known domain: sports, and the availability of robust event detectors based on content analysis, it is possible to resolve many of the outstanding issues posed in [10].

In [6] the MHI was used directly as a representation of motion in a single still of a game of snooker. The MHI is a projection of image material from surrounding frames into a current frame, at locations of high displaced frame difference. Used in its original form it results in a the creation of a mask that covers the entire 2D region swept out by a moving object (c.f. making angels in snow). Modified to be used as an effect (as in [6]) it is an implicit technique that results in a comet tail around moving objects. In [6] a still camera was assumed, however this work takes the idea further by coping with moving cameras and arbitrary shape objects.

One key contribution of this work is to connect content based analysis with the creation of a *salient still*. The other key contribution is to avoid the need for explicit segmentation and instead rely on the content based analysis in such a way that *implicit* segmentation occurs. The pictures presented here consider Tennis as an example of the process.

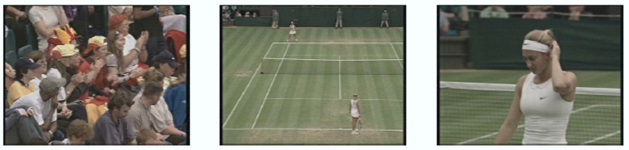


Fig. 2. *Three images showing typical shots in tennis footage. The full court view shot (centre) contains the main semantic game atom: the act of hitting the ball.*

Consider for the moment that a sementically relevant event in the video stream has been detected at a particular frame. In tennis, the basic game atom is the instant that a player makes contact with the ball, and that event will be used to demonstrate the idea proposed here. Define this event the *impact event*. Figure 2 shows a frame from three typical shots in grass court Tennis. The centre shot showing the full view of the court the shot of significance and is the focus in this paper. That shot contains all the *impact events*.

Because the semantics of the detected event are known a-priori it is possible to exploit the fact that only moving objects in this atom are relevant. The idea therefore is to create a single frame that represents this motion in some sense. The obvious approach is to register moving objects in frames before and after the instant of interest onto the current frame, and superimpose those objects in a single frame, hence *cartoonising* the motion in a sense.

2.1. Global Motion Compensation

Define the frame containing the impact event as I_n. It is necessary to compensate surrounding frames for camera motion in the view. This can be done with a variety of existing estimators [12, 8]. Here a bilinear global motion model is used such that

$$I_n(h,k) = \\ I_{n-1}(a_0h + a_1k + a_2hk + dx, b_0h + b_1k + b_2hk + dy)$$

where the pixel intensity at position (h,k) in frame n is $I_n(h,k)$. $\{a_j, b_j\}$ are the six bilinear transformation parameters, with dx, dy the translation. In practice this model was the best computational compromise between purely affine and perspective transformations that would allow compensation of the slight camera rotation. Using integral projections [4] at the original image resolution a reasonably robust estimate of dx, dy can be generated. A version of the estimator presented in [12] is then used to estimate the bilinear parameters, by exploiting a coarse to fine refinement strategy with 2 levels of image downsampling. Each frame $I_{n-1}, I_{n-2}, I_{n-3}, \ldots, I_{n-p}$ is registered with respect

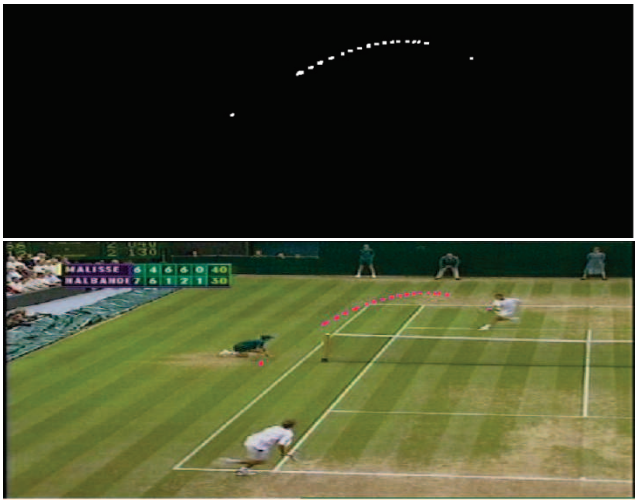


Fig. 3. *Top: Giving the appearance of ball tracking by rough segmentation using displaced frame difference and colour. Ball masks from 18 frames have been extracted. Bottom: Superposition of each mask on frame I_n in colour.*

to I_n separately in order to achieve the best final effect. Furthermore, experiments with creating the Motion Based Keyframe using both next and previous frames around I_n indicate that users prefer the Motion Keyframe to include a representation using previous frames only. For tennis $p = 12$ frames seems sufficient.

2.2. Implicit motion representation

The motion of three objects must be represented in the Key Frame. The ball and the two players. The ball in tennis moves fast enough that the difference between global motion compensated images is significant at ball locations. Define the displaced frame difference between two frames as follows

$$\Delta_{q,q-j}(h,k) = I_q(h,k) - I'_{q-j}(h,k) \tag{1}$$

$$\Delta_{q,q+j}(h,k) = I_q(h,k) - I'_{q+j}(h,k) \tag{2}$$

where $I'_{q-j}(\cdot,\cdot)$ is the image $q-j$ motion compensated with reference to the image at q with the global motion estimated previously. At locations where the foreground has moved, *both* the forward and backward differences are large. Hence a binary mask $L_q(h,k)$ can be created denoting with 1 the presence of a moving object and 0 otherwise as follows

$$L_q(h,k) = \begin{cases} 1 \ \text{if} \ (|\Delta_{q,q-j}(h,k)| > T) \ \text{AND} \ (|\Delta_{q,q+j}(h,k)| > T) \\ 0 \ \ \text{Otherwise} \end{cases}$$

$L_q(h,k)$ is a good indicator of the location of foreground when the motion is large, hence $j = 2$ for creation of the pictures shown later. To remove noise a morphological closing operation is performed. However, L_q covers both player and ball location. By retaining only closed regions larger than 2×2 pixels and less than 10×10 pixels, the ball regions can be selected. These regions are then marked on frame n by changing the colour of the pixels there. Alternate representations could enlarge these regions and superimpose them on the Motion Key frame. Figure 3 shows the effect created. Note that the ball track is clearly visible even though *no explicit object tracking* has been performed.

2.2.1. Player delineation

Representing the motion of the player is slightly more difficult. Simply superimposing the player mask into the current frame is confusing. Instead, it is possible to superimpose the outline of each player in previous frames into the current frame n. Unfortunately, the caption can also be included into the player mask because it often has some apparent motion. Explicit tracking of the tennis players allows a region of interest to be delineated around each player and hence the creation of player masks for manipulation. While the methods presented in [6, 13] using particle filters are viable, a lower cost technique works well here.

There are two features that can be used to locate players in view: motion, indicated by the energy of Δ above and colour. For wimbledon for instance, the tennis players wear white. A rough colour mask can be generated by detecting pixels which are loosely white for instance. Assuming that players generally keep to the top or bottom half of the view it is possible to integrate motion and colour information vertically and observe that the result contains peaks corresponding to possible horizontal locations of the player. Only the horizontal position is extracted since in general all that is required is to delineate a vertical region of interest in the image for superimposing. By selecting the 5 largest peaks as candidate locations, the viterbi algorithm can be used to track the best horizontal location path of the player through the set of frames.

The features are defined as follows

$$S_q^+(h) = \sum_k \Delta_{q,q+1}(h,k)$$

$$S_q^-(h) = \sum_k \Delta_{q,q-1}(h,k)$$

$$C_q(h) = \sum_k (I_q^R(h,k) > 30) + (|I_q^G(h,k)| > 30) + (|I_q^B(h,k)| > 30)$$

where I^R, I^G, I^B are the Red, Green and Blue components of colour at the relevant pixel location. The combined feature is then $p_q(h) \propto 0.5(S_q^+(h) + S_q^-(h)) + C_q(h)$. $p_q(h)$

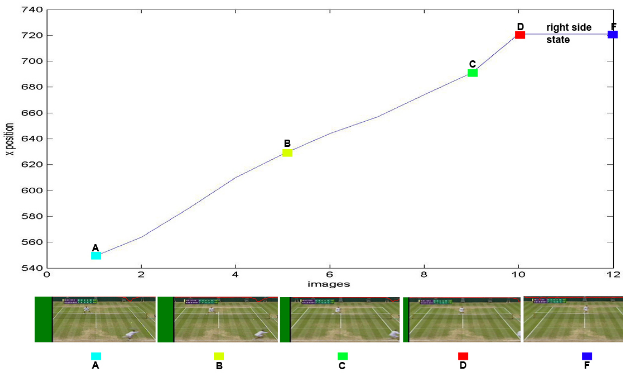


Fig. 4. *Top: a plot showing the estimated horizontal position of a player versus frame. Bottom: the corresponding frames. Note that the right side state is called into play as the player moves off the right of the screen.*

is normalised to one, and can be considered as a measure of probability that the object exists at horizontal location h in the image. This is the data likelihood.

The transition probability connecting object position between frames is defined as follows

$$p(x_{q+1}|x_q) \propto \exp\left(\frac{-(x_{q+1}-x_q)^2}{2\sigma_x^2}\right) \tag{3}$$

where x_q is the player position in frame q.

In order to cope with the rare situation in which the player moves off the left and right hand side of the field of view, two outlier states are introduced denoted to have position $x = -1$ and $x = M$ respectively. M is the horizontal resolution of the image. The associated likelihood is inversely proportional to the number of detected *white* pixels. Thus, if there are no white pixels, implying that there could be no player in view, then these off-screen states are more likely. The likelihood is as follows

$$p_q(h) = \frac{1}{5}\exp - \left(\frac{\sum_j C_q(j)}{2\sigma_s^2}\right) \tag{4}$$

where $\sigma_s^2 = 30$ for the pictures generated here. There is some experimental choice in setting this relationship, but the notion is to allow the off-screen states to compete on the same basis for the track as the on-screen positions.

Using this configuration the viterbi algorithm can be used to track the player horizontally. Figure 4 shows how the tracker works across 12 frames of a full court view.

2.2.2. Onion Skinning

Given an estimate of player location through the registered group of frames, a region of interest for the mask $L_q(h, k)$ can be generated. Within this region of interest, significant gradients are marked in colour on the target frame n. By exploiting the position information, outlines can be drawn only when the player object no longer overlaps with its "skin" from previous frames. To render a sense of time on the pictures, the colour is changed depending on how far in the past the skin originated. Figure 5 shows the created effect at one impact event.

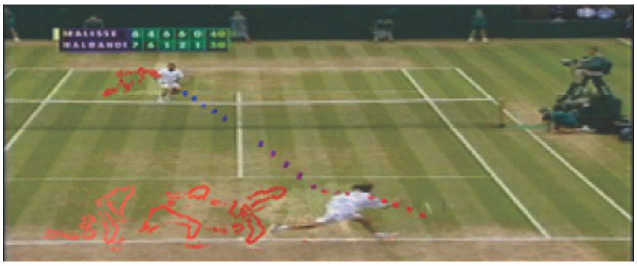


Fig. 5. *Final representation of player and ball. Note as the effect is to superimpose time history on a single frame, this implicit method does not occlude the ball track with the player in the current view.*

3. CONTENT ANALYSIS

In Tennis, detection of the impact event can be achieved by combining an audio detector with a view classifier. A visual feature can be created by examining implicit scene geometry. That geometry is used to classify shots containing a full view of the court. The audio track gives information about the instant that a ball is hit when such views are identified.

3.1. Implicit Geometry

The second moment of the hough transform of an edge map of each image is computed for each image [?]. This measure, noted $\mathbf{x}_v$, is used to detect frames showing a main view of the court where its value remains constant. Hough space contains peaks corresponding to image lines, and in full court views, for example as shown in the centre image of figure 2, there is strong geometry and hence Hough space is populated with few compact clusters. The orientation of these peaks therefore changes with geometry of the input image which in turn is affected by the court view. As the full view of the court is dominated by the physical, rectangular court structure the feature works well to discriminate it without the need of any 3D information (as used in [14]). Figure 6 shows the plateaus in moment measure that correspond to the court views in an example of tennis footage. The likelihood of this feature given a main view of the court can be expressed as

$$\mathcal{P}(\mathbf{x}_v|\text{Full view}) \propto \exp\left[\frac{-(\mathbf{x}_v-\mu_v)^2}{2\ \sigma_v^2}\right] \tag{5}$$

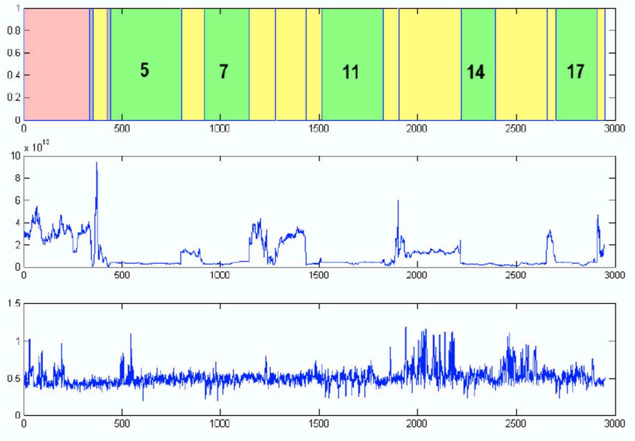


Fig. 6. *Top: ground truth over the sequence (rose: crowd; green: full view of the court, yellow: close up of the players, grey: dissolve; black: cut), Middle: Moment of the hough space, Bottom: DFFS computed using audio features.*

where the mean μ_v and variance σ_v^2 are estimated from training sequences.

3.2. Frame level audio feature

Since the racket hit is a short sound between 10 to 20 ms long, a spectrogram of the audio track using a 40 ms window (duration of a frame in the video) is sufficient to express the relevant temporal resolution. The power spectrum of this Fourier transform, normalised by its energy, is then computed for each window and corresponds to the audio features $\mathbf{x}_a$.

Eigenspace representation. K audio features corresponding to racket hits are collected. A Principal Component Analysis (PCA) is then performed over this training database. J eigenvectors corresponding to the J highest eigenvalues are retained to span the eigenspace F.

Distance from the feature space. The similarity of an unknown observation $\mathbf{x}_a$ with the training cloud, is estimated by computing the distance between $\mathbf{x}_a$ and the eigenspace F. This *Distance From Feature Space* (DFFS) is defined as [11]:

$$\text{dffs}(\mathbf{x}_a) = \|\mathbf{x}_a - \mu_a - \text{UU}^{\text{T}}(\mathbf{x}_a - \mu_a)\| \quad (6)$$

where μ_a is the mean of the audio features, and U is the matrix collecting the J eigenvectors computed in the learning step with PCA.

Likelihood of having a Racket hit. Assuming a uniform distribution over the eigenspace F, the likelihood of having a racket hit can be approximated [11] using the likelihood of the reconstruction error :

$$\mathcal{P}(\mathbf{x}_a|\text{Racket hit}) \propto \exp\left[\frac{-(\text{dffs}(\mathbf{x}_a))^2}{2\,\sigma_a^2}\right] \quad (7)$$

σ_a^2 is estimated using the mean value of the eigenvalues $\{\lambda_j\}_{j>J}$ in $F^{\perp}$ [11].

3.3. Video features at shot level

Frame level video features are processed to generate shot level features. These features allow access to higher level content information, in classifying shots as rallies R or not $\overline{R}$.

Shot level visual feature. The mean likelihood of the moment feature over a shot is used to represent the visual content of a shot $\mathbf{x}_v^s$:

$$\mathbf{x}_v^s = \mathbf{E}_{\mathbf{x}_v \in s}\left[\left(\frac{\mathbf{x}_v - \mu_v}{\sigma_v}\right)^2\right] \quad (8)$$

This feature is independent of the length of the shot and its likelihood is expressed as

$$\mathcal{P}(\mathbf{x}_v^s|s = R) \propto \exp - \left[\frac{\mathbf{x}_v^s}{2}\right] \quad (9)$$

Shot level audio feature. Shot audio content is estimated as the minimum of the similarity measure DFFS computed over the shot $\mathbf{x}_a^s$:

$$\mathbf{x}_a^s = \min_{\mathbf{x}_a \in s}\left\{\left(\frac{\text{dffs}(\mathbf{x}_a)}{\sigma_a}\right)^2\right\}$$

The likelihood is then:

$$\mathcal{P}(\mathbf{x}_a^s|s = R) \propto \exp - \left[\frac{\mathbf{x}_a^s}{2}\right] \quad (10)$$

Fusion of audio and visual information. Assuming the independence of audio and visual data, the likelihood using both audio and visual features has simply been computed by:

$$\mathcal{P}(\mathbf{x}_a^s, \mathbf{x}_v^s|s = R) = \mathcal{P}(\mathbf{x}_a^s|s = R) \times \mathcal{P}(\mathbf{x}_v^s|s = R) \quad (11)$$

Figure 6 shows the ground truth over one test sequence (of 2500 frames): shots $\{5, 7, 11, 14, 17\}$ belong to the class of interest R. The middle plot shows the second moment of the hough transform. Plateaus correspond to full court views but also at player close ups. Therefore false alarms may appear when using only visual features. The bottom plot presents the similarity (DFFS) computed over the sequence. Low values indicate high probability of racket hits. Visually, it can be seen that while both detectors separately can show false alarms, together they achieve very good performance despite their simplicity. See [5] for further experimental evaluation of this detector.

4. PICTURES AND APPLICATIONS

Figure 7 shows three events rendered with Motion Keyframes at time instants detected by the impact detector discussed above. Viewers preferred skins to preceed the event rather than follow it. A zoom on a single frame representation is shown in figure 5. As can be seen, player colour fades with time, while only outlines are represented.

These frames can also be used to create a content sampled sequence. In such a sequence a new frame is transmitted only when an interesting event occurs, and at that time, the Motion Keyframe can be transmitted instead of the real frame. This would convey a sense of motion to the viewer who would then be viewing essentially stop action footage. However, experiments show that viewers prefer some image sequence to be shown even though nothing of semantic relevance may be seen. To resolve this problem any single frame within a 1 sec window (for example) is shown if there is no impact event. If an impact event does occur in that window the motion keyframe is used. Experiments show that such content aware low frame rate sequences are easier to understand than non-content aware sequences. It is difficult to convey the effect of this kind of sequence in the format of this publication. The reader is directed to view the video examples at
`www.mee.tcd.ie/∼sigmedia/publications/cbmi05/`
to gain an appreciation of content aware sampling.

In practice content aware streaming such as this can be more appealing if the objects appear to move against a static background. Otherwise, each new Motion Keyframe can appear to jump by a large amount horizontally. This requires rendering Motion Keyframes against the entire backdrop e.g. the entire court in tennis. This is possible by mosaicing frames (see fig. 8) across the shot. Given that the content analysis engine can reliably extract the sequence frames for which a mosaic is feasible, it becomes easier to process a continuous broadcast.

5. FINAL COMMENTS

This paper has presented the notion that content based analysis can be used to resolve many issues in *hard* image processing and computer vision problems by allowing the confident application of contextual information. Further, it has presented the idea of Motion Keyframes that attempt to represent the motion across several frames in a single image. The combination of content based analysis and content representation allows sequences to be viewed *by content* rather than by time. In addition, the Motion Keyframe can better represent content when used as a menu for summary purposes for instance. While it must be acknowledged that the effects shown here may not be of cinematic quality, there is wide scope for adding to the current human-user centric focus of content based analysis by acknowledging that it also could be used to control the low level processing of media.

6. REFERENCES

[1] J. Assfalg, M. Bertini, C. Colombo, and A. Del Bimbo. Semantic annotation of sports videos. *IEEE Multimedia*, 9(2), Apr/Jun 2002.

[2] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3), Mar 2000.

[3] S.F. Chang, D. Zhong, and R. Kumar. Real-time content-based adaptive streaming of sports video. In *IEEE Workshop on Content-Based Access to Video/Image Library*, pages 139–146, Dec 2001.

[4] A. Crawford, H. Denman, F. Kelly, F. Pitié, and A. Kokaram. Gradient based dominant motion estimation with integral projections for real time video stabilisation. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 3371–3374, October 2004.

[5] R. Dahyot, A. C. Kokaram, N. Rea, and H. Denman. Joint audio visual retrieval for tennis broadcasts. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2003.

[6] H. Denman, N. Rea, and A. Kokaram. Content-based analysis for video from snooker broadcasts. *Journal of Computer Vision and Image Understanding, Special Issue on Video Retrieval and Summarization*, 92:141–306, November/December 2003.

[7] R.D. Dony, J.W. Mateer, and J.A. Robinson. Automated reverse storyboarding. In *IEE 1st European Conference on Visual Media Production*, pages 193–202, March 2004.

[8] F. Dufaux and J. Konrad. Efficient, robust and fast global motion estimation for video coding. *IEEE Transactions on Image Processing*, 9:497–501, 2000.

[9] A. Ekin, A. M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transaction on Image Processing*, 12(7):796–807, July 2003.

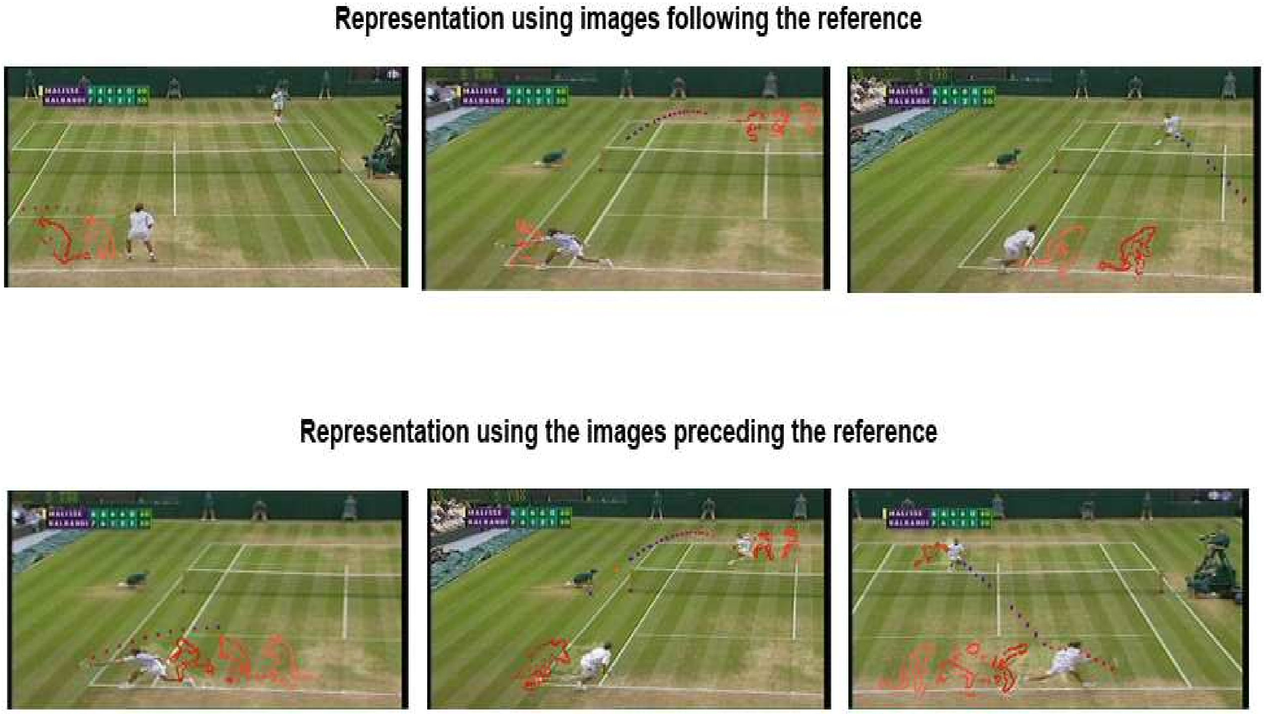
[10] M. Massey and W. Bender. Salient stills: Process and practice. *IBM Systems Journal*, 35(3,4), 1996.

[11] B. Moghaddam and A. Pentland. Probabilistic visual learning for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, Juillet 1997.

[12] J-M. Odobez and P. Bouthémy. Robust multiresolution estimation of parametric motion models. *Journal of visual communication and image representation*, 6:348–365, 1995.

[13] N. Rea, R. Dahyot, and A. Kokaram. Modeling high level structure in sports with motion driven hmms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 621–624, May 2004.

[14] G. Sudhir, J. C. M. Lee, and A. K. Jain. Automatic classification of tennis video for high-level content-based retrieval. In *IEEE International Workshop on Content-Based Access of Image and Video Databases*, pages 81–90, Jan 1998.



Representation using images following the reference

Representation using the images preceding the reference

Fig. 7. *Representation by superimposing previous or following frames. Viewers prefer previous frames as* skins.

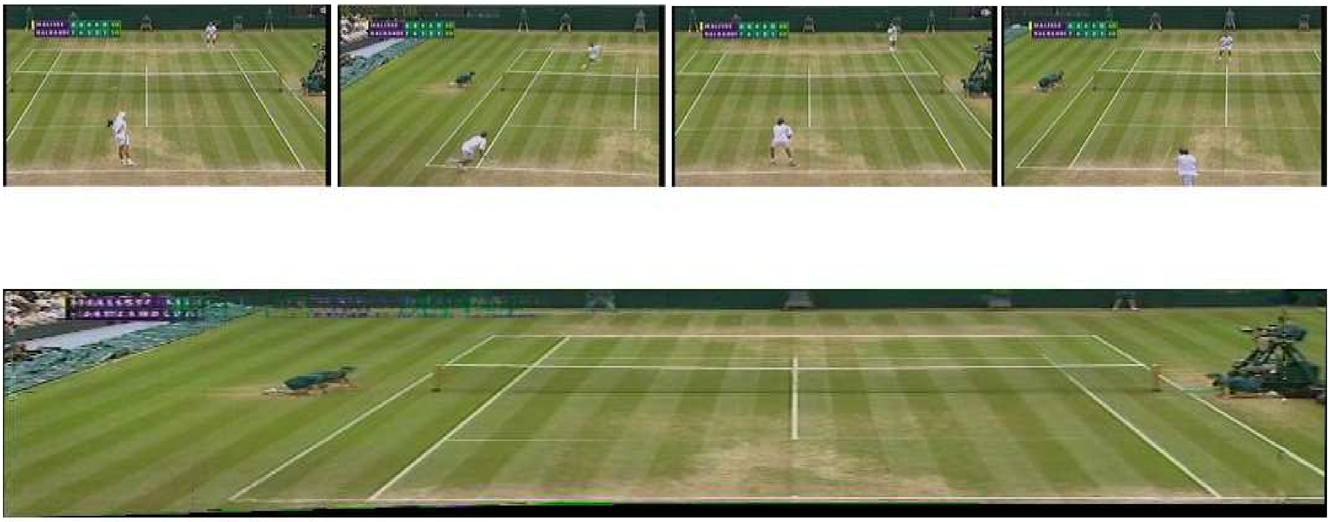


Fig. 8. *Creating a mosaic across the full court view shot. Top 4 frames out of 18 used for the mosaic shown at the bottom.*

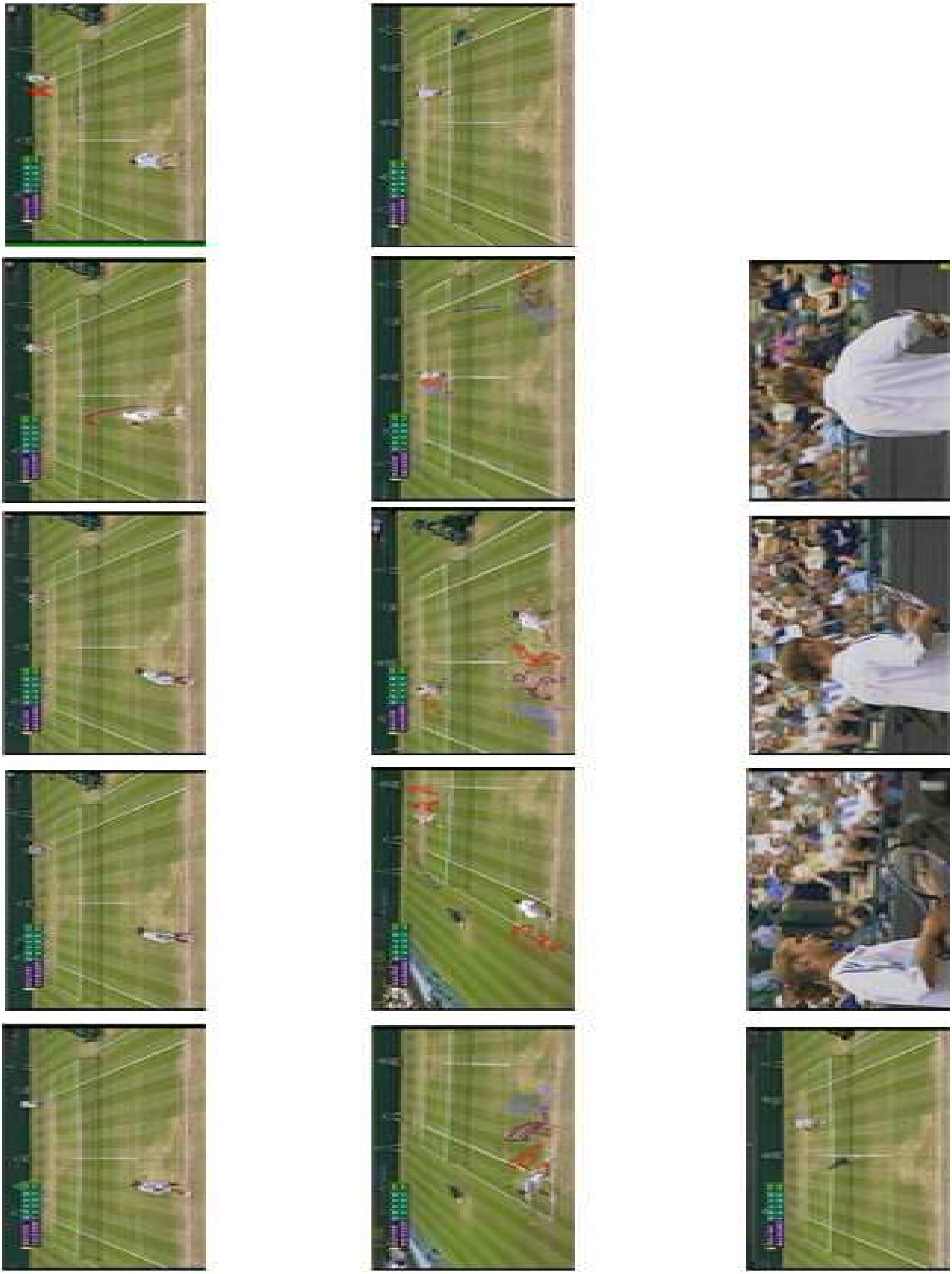


Fig. 9. Summarization with motion keyframes.