Determining Shot Assonance/Dissonance via Salience Maps and the Match Frame Principle of Continuity Editing

Robert Turetsky and Xanadu Halkias
Columbia University Department of Electrical Engineering
{rob, xanadu}@ee.columbia.edu

Abstract

One of the ultimate challenges of computer vision is in video semantic understanding. Many efforts at detecting events in video have focused on structured sequences such as sports or news broadcasts. However even in seemingly freeform media such as feature films, there exists inherent structure and established production codes. Over the last century, film theorists have developed the principles of continuity editing. One tenet of continuity editing is known as match framing*: in order for a shot boundary to appear seamless, the viewer's focus of attention should not have to move very far from one shot to the next. Filmmakers will generally adhere to the continuity editing guidelines in order for audiences to maintain their suspension of disbelief. Often times, however, prudent violations of continuity can jar the viewer, for example during action scenes or moments of high intensity. By detecting violations of the continuity editing principles, it is possible to locate portions of a film that the filmmaker is interested in portraying as different from the rest of the film.*

We have developed a method for automatically detecting violations of the match framing principle that fuses film theory, psychophysical modeling, image morphology and pattern recognition. First, shot detection is performed on the entire film. Next, we compute the saliency map on a frame before and after the shot boundary. We then treat each saliency map as a distribution, and estimate a 3-component Gaussian Mixture Model of the salient peaks. Finally, by comparing distributions we are able to estimate how active the viewer's eye will need to be from one shot to the next. Experiments demonstrate a correlation between match frame violations and plot in a small corpus of full-length movies.

1. Introduction

One of the most important aspects when dealing with the manipulation of video is the extraction of information that would facilitate digital exploitation such as indexing/retrieval, summarization, compression and transcoding. Computer Vision looks for structural aspects of an image in order to incorporate an understanding on the works of the human visual system.

However, in recent years, most of the existing literature on video analysis deals with highly standardized video sequences such as sports or news broadcasts [4,13,15]. This is mostly due to the immediate commercial needs of the above-mentioned formats and also because their structure provides an intuitive guide on the desired processing steps.

So far, there is a small specimen of work done on full-length feature movies [5,8], most of which is only in a very restricted framework [13] e.g. trailers, portions of the movie. This is due to their extensive computational and storage requirements.

In this paper we work with full movies and incorporate the knowledge from film theory in order to extract features that represent the underlying structure of narrative filmmaking as presented in the remaining part of this section.

Since the Russian formalists began meeting in film clubs at the turn of the 20th century, film theorists have attempted to codify the principles of audiovisual storytelling. In the 1920s, Lev Kuleshov [7,11] demonstrated that editing, the juxtaposition of shots in time, could create meaning in two otherwise disparate shots. Over time, the rules of *continuity editing* evolved. Continuity editing are the guidelines with which a filmmaker can place the camera and the editor can splice two shots together in order to mask the effect of the camera. There are six important rules [1] that are paramount in achieving the desired result: The 180-degree rule, the 30-degree rule, cutting-on-action, match framing, matching eye lines and script continuity. By adhering to continuity, a filmmaker can maintain suspension of disbelief – the viewer will feel that they are watching reality, when in fact what they are witnessing is an extended stream of carefully ordered shots. In this paper, we focus on the rule of match framing.

1-4244-0028-7/06/$20.00 ©2006 IEEE

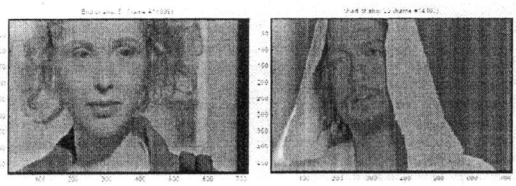

Figure 1: A match frame, on the boundary between shots 54-55 in *Killing Zoe*. The center of focus (the actors' eyes) is in the same location from shot 54 to shot 55.

A shot boundary is a drastic change in the visual content of the film. If a spectator witnesses a jarring cut, the fact that they are watching a film will instantly call attention to itself. The principle of *match framing* [1] serves to minimize this effect, as seen in Fig. 1. It states that a fluid shot boundary will take place if the center of attention in the first shot will be at or near the center of attention in the second frame. When a human is speaking in the frame, the viewer will focus their attention on the speaker's eyes. Note that the salient focal point can move during the course of a single shot.

Filmmakers often choose to violate continuity in order to create a brash visual effect. Violations of the match framing principle, for example, can engage the viewer to be a more active participant in the film, as their focus of attention will need to shift frequently. This is appropriate, for example during action sequences, or periods of harsh and intense emotions. Automatically detecting match-framing violations will lead to the uncovering of these types of sequences within a produced motion picture, which would be useful for automatic summarization.

The remainder of this paper will proceed as follows. First, we detail our method for detecting match framing. We present a qualitative analysis of our method in Section 4, and concluding remarks in Section 5.

2. Automatic Detection of Match Framing

As mentioned above, continuity editing defines the principles of masking edits. We defined a match frame as the process when the center of attention between two adjacent shots is in the same area of the frame. In the remaining section, we will detail our system for automatically detecting match frames and violations of the match framing principle.

2.1 System Overview

Our system for detecting match frames is as seen in Fig. 2. First, we perform shot detection on the entire film. Second, we create a saliency map on shots two frames before and two frames after the shot boundary. By treating each saliency map as a distribution, we can estimate a GMM on the distribution. Finally, we can

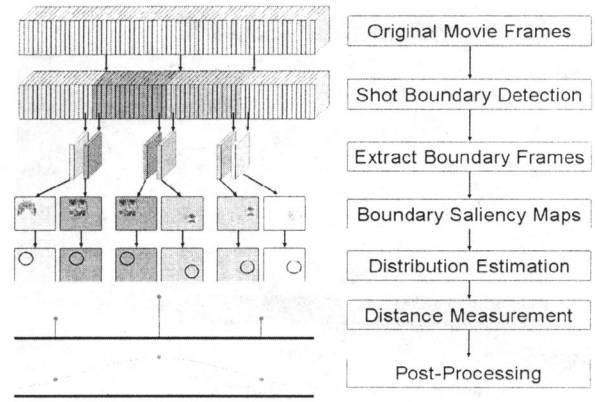

Figure 2: System Overview

obtain an estimate of the distance the viewers' eye must travel on the cut by comparing distributions. Each component of the system will now be described.

2.2 Shot Boundary Detection

We detect shot boundaries using the edge method [9]. Although edges are computationally intensive to compute on an entire film, we believe that this method is appropriate because it is robust to real-world shots, which often do not vary much in color histogram, and to slow transitions.

2.2.1 Pre-processing

First the individual frames of the film are pre-processed to remove the black border on the top and bottom of the frames that are inserted because of letterboxing as seen in Fig. 3. This is because the borders introduce unintended edges and color gradients that "confuse" the shot and salience detectors. We can find the frame borders by summing the intensities of $N=40$ randomly selected frames. For each row of this sum, we compute the number of pixels that are in the top 1.5% of the intensity range (e.g. 256*40, where black=1, white=256). The first row with half the pixels above threshold is the top border, and the next row with half the pixels above the threshold is the bottom border. The above is shown in Eq. 1.

$$S(x,y) = \sum_{i=1}^{N} I_i(x,y)$$

$$\Phi(x,y) = 1_{S(x,y)>th}$$

$$\min_x \; s.t. \; \sum_x \Phi(x,y) > \frac{y_{\max}}{2} \qquad (1)$$

$S(x,y)$ is the superimposed image of the $N=40$ random frames, $\Phi(x,y)$ is an indicator matrix and y_{max} is the number of columns in the frame. The threshold

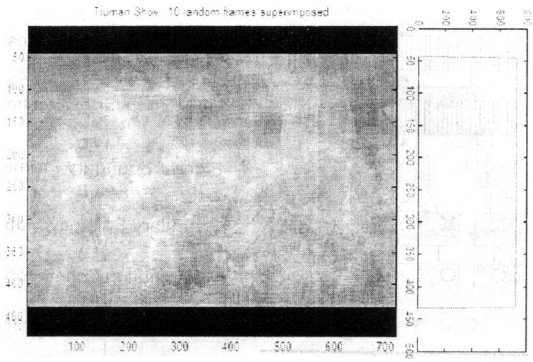

Figure 3: Frame border computation. 10 superimposed random frames from *The Truman Show*.

th is 4*40, where 4 is approximately 1.5% of the grayscale. The same procedure can be implemented to extract the right left borders if they exist.

2.2.2 Shot Boundary Detection

Our shot boundary detection system follows [9]. Edges are computed for two adjacent frames *i, j* using the Canny method [2]. Frame *j* is dilated by a 7-point structuring element, and inverted. We compute ECR^{out}, the number of edges leaving the frame, as the sum of the logical *and* of the edges of frame *i* and the dilated/inverted edges of frame *j*, normalized by the number of edge pixels in frame *i*. ECR^{in} is likewise computed as the logical *and* of frame *j*'s edges and the dilated/inverted edges of frame *i*, normalized by the number of edge pixels in frame *j*. The likelihood that this is a shot boundary is taken as $min(ECR^{in}, ECR^{out})$. The above is seen in Eq. 2-4.

$$ECR^{out} = \frac{\sum E_{I_i(x,y)} \cap E_{I_j(x,y)\oplus B}}{\sum E_{I_i(x,y)}} \quad (2)$$

$$ECR^{in} = \frac{\sum E_{I_i(x,y)\oplus B} \cap E_{I_j(x,y)}}{\sum E_{I_j(x,y)}} \quad (3)$$

$$\ell = \min(ECR^{in}, ECR^{out}) \quad (4)$$

Peak picking proved to be the most difficult part of the algorithm. In order to remove false minima and smooth out fluctuations within shots with a lot of movement, we subtract the mean of each point around an $L=15$ point normalized Hamming window, $w[k]$. This corresponds to approximately 0.5 sec of screen time. Before subtracting the mean, we add a small constant, ζ to it (our empirical study led to a value of $\zeta = 0.12$) in order to remove local maxima originating from noise that lie above the average, Eq. 5. The resultant curve is half-wave rectified, and only peaks whose original shot boundary likelihood was greater than a global threshold of 0.4 are kept.

$$\max(0, \ell[n] - \sum_{k=n-\lfloor L/2 \rfloor}^{n+\lfloor L/2 \rfloor} \ell[k]w[k] + \zeta) \quad (5)$$

In our evaluation, we properly detected all 889 shot boundaries in *Killing Zoe*, however we also had 87 false positives. The false positives generally arise from rapid camera movement. False positives which would have arisen from camera panning are neutralized by automatically aligning frames before dilation. As our final results depend only on global values as opposed to individual shots, a nonzero error in shot boundary computation can be tolerated.

For each shot boundary, we need to compare a frame from the end of the previous shot and the beginning of the next shot in order to determine if frame matching has been met. We choose two frames before and two frames after the shot boundary, because the DVD MPEG-2 stream sometimes has interlacing errors around shot boundaries. The saliency map is described in the next section.

2.2.3 Saliency Map

Now that we have the shot boundary estimations, we can compare frames from both sides of the shot boundaries. In order to highlight perceptually salient features of the frames in question, we create a saliency map. The saliency map was developed by Itti and Koch [6], and fuses saliency calculations in color, intensity and orientation at various scales. The process of creating a saliency map is illustrated in Figure 5.

The operating theory behind the saliency map is to extract the foreground from the background by down-sampling the region of the image, then up sampling it to the original resolution in order to compare the image with its down-sampled version. The idea behind this is that the fine details will be left out in the lower levels of the pyramid, and if they exist, differences between foreground and background can be considered "salient."

The first area of processing is the color/intensity. We take the original full-size movie frames and create color maps using broadly tuned channels, Eq. 6-10:

$$I = (r + g + b)/3 \quad (6)$$

$$R = r - (g+b)/2 \quad (7)$$

$$G = g - (r+b)/2 \quad (8)$$

$$B = b - (r+g)/2 \quad (9)$$

$$Y = (r+g)/2 - |r-g|/2 - b \quad (10)$$

For each color/intensity map, we create a Gaussian pyramid by blurring each map with a Gaussian filter and then down-sampling by a factor of two. We look

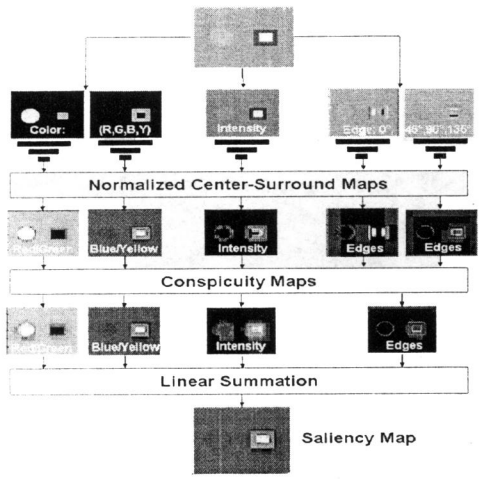

Figure 4: Extraction of Saliency Maps

for salient features using center-surround maps. Center-surround maps simulate color gradient cells in the visual cortex, which look for clashes between colors that are perceptually different (red/green, blue/yellow, light/dark). We create 6 center-surround maps for intensity and 12 for color Eq. 11-13:

$$I(c,s) = |I(c)\Theta I(s)| \quad (11)$$
$$RG(c,s) = |(R(c)-G(c))\Theta(G(s)-R(s))| \quad (12)$$
$$BY(c,s) = |(B(c)-Y(c))\Theta(B(s)-Y(s))| \quad (13)$$

Where Θ defines a point-to-point difference and c, s are scales such that $c=\{2,3,4\}$, $\delta=\{3,4\}$, $s=c+\delta$. Each map is normalized by multiplying it with $(M-m)^2$ where M = global max of the map, and m be the mean of the map's local maxima without the global max.

In order to create orientation maps, we first create a Gabor pyramid, by utilizing 2-D Gabor filters [3]. In order to create center-surround maps, we use Eq. 14:

$$O(c,s,\theta) = |O(c,\theta)\Theta O(s,\theta)| \quad (14)$$

Where c and s are defined as above, and $\theta=\{0°, 45°, 90°, 135°\}$ is the orientation of the filter. Additionally, each orientation map is normalized as above. Since we have 6 maps per orientation and 4 orientations, this gives us 24 maps for orientation.

We aggregate all of the maps for color, intensity and orientation respectively by creating three "conspicuity maps." These are simply re-sampling each normalized center-surround map to pyramid level 4 and summing. Adding together all 3 conspicuity maps creates the final saliency map, Fig. 4.

The saliency map provides an indication of where the viewer's focus of attention is expected to be. In general, this occurs at places where the foreground differs from the background in color or intensity, and in areas of high textual detail.

The filmmaker's job is to guide the viewer to what is important in a given frame. Thus it is assumed that the filmmaker will compose each shot in such a way that the most perceptually salient regions of the frame are where the filmmakers would want the viewers' attention to focus. Thus, the saliency maps provide a good approximation to the intentions in the composition of the shot.

The last stage is to compute the distance between frames on both sides of the shot boundaries.

2.2.4 Estimating Shot Dissonance

Once we have the saliency map, we are able to estimate the dissonance (distance) between frames of the prior and current shot. We begin by computing the saliency maps for the frames two ahead and behind a shot boundary estimate.

Invariably, there are between one and four distinct regions of importance in the saliency map. In general, one or two of these are "true" peaks in salience, while the remaining peaks are "false positives." The natural thing to do is to model the saliency maps as Gaussian Mixture Models, and compare the frames by comparing the distributions of the two models.

We estimate a 3-component GMM for each of the two frames using EM. From there, we choose the two highest weighted Gaussians from each frame, and compute the Euclidean distances between the two components in each frame. Supposing we have two means μ_{l1}, μ_{l2} from the previous saliency map and μ_{n1}, μ_{n2} of the current saliency map, the dissonance estimate of shot i is, Eq. 15-17:

$$d_1 = |\mu_{l1} - \mu_{n1}| + |\mu_{l2} - \mu_{n2}| \quad (15)$$
$$d_2 = |\mu_{l2} - \mu_{n1}| + |\mu_{l1} - \mu_{n2}| \quad (16)$$
$$\min(d_1, d_2) \quad (17)$$

A large number means a high degree of dissonance between shots. A small number means it is likely that we have a matched frame.

2.2.5 Post-Processing

Film directors can create a feeling of intensity by tension-relaxation, a technique of alternating between two extremes of an audiovisual feature. Within the high-intensity segments, it makes sense that match framing violations would not be continuous, as excitement can be created by varying the distance between shots within a scene. This suggests that patterns can be found by lowpass filtering the shot assonance/dissonance curve.

Figure 5A tracks the shot assonance/dissonance in *Killing Zoe*. The peaks that rise above the noise floor correspond to violations of match framing. Figure 5B shows this curve with the mean subtracted, smoothed with a 21-point moving average filter, and half-wave rectified. This is the *MFVC* – match frame violation

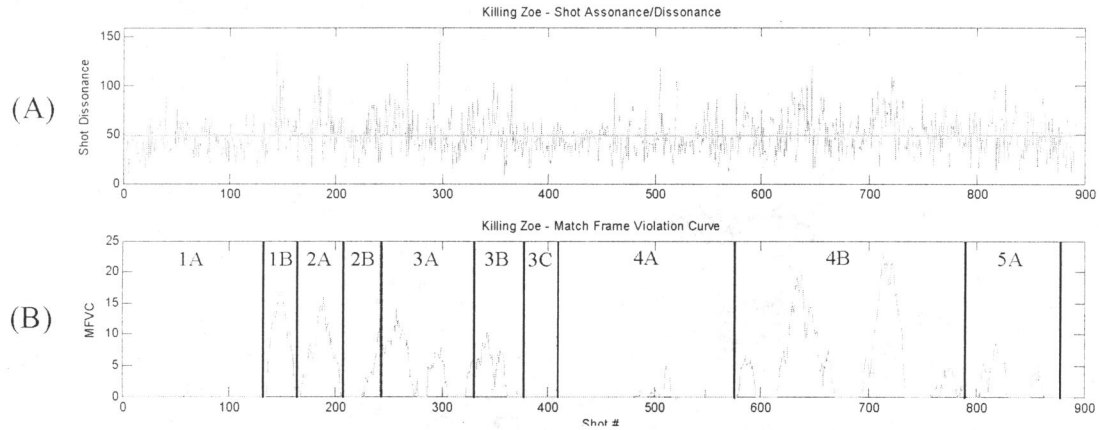

Figure 5: Shot Assonance/Dissonance (A, top) and Match Frame Violation Curve (B, bottom)

curve. Note the correspondences between scenes and their predicted shot assonance/dissonance values.

3. Qualitative Analysis

In order to assess the effectiveness of this algorithm, we must perform qualitative analysis. This method was never meant to discover all intense scenes in a film, but instead to work in tandem with a battery of other tests. Thus a quantitative analysis of this method will only be able to be performed within the confines of a feature selection experiment.

The film *Killing Zoe* has five main sections:

1) *The arrival*: Zed arrives in Paris, meets Zoe. They fall in love (1A, shots 1-132). Eric arrives and throws Zoe out of the hotel room (1B, shots 132-164).
2) *The plan*: Zed and Eric meet up with the gang (2A, shots 165-206). They plan the big bank heist (2B, shots 207-242), which seems like a good idea except that the job is tomorrow.
3) *The party*: The gang goes out to party, which is crazy (3A, shots 243-330) to overwhelming (3B, shots 330-377), to a sickening haze (3C, shots 378-409).
4) *The heist*: Eric, Zed and the others attempt to rob a bank. At first, they are in control (4A, shots 410-575), but then everything goes wrong (4B, shots 576-790).
5) *The resolution*: Zed, Zoe and Eric have a final confrontation (5A, shots 791-878). Zoe saves Zed and they leave together (5B, shots 878-889).

It would make sense that in section 1A, 2A and 4A and 5B would have a high shot assonance. Theory would dictate that there would be a high incidence of match frame violations in 1B, 3A/B, 4B and 5A.

In the "boy-meets-girl" scene in 1A, the MFVC is near zero. When Eric arrives and throws her out in 1B, the MFVC will peak to show the heightened action. The MFVC is high during 2A, which matches the enthusiasm that Zed has when meeting the gang of thieves. In 2B, the MFVC starts at zero, possibly to symbolize the "coolness" involved in the planning of the heist, but then it rises when it is evident that they are in over their heads. We have alternating highs and lows during the party, which closely mirrors its tumultuous nature.

In 4A, the bank robbers are in control – everything is going smoothly. Thus, we have a lot of match frames. There are a few peaks present in the assonance/dissonance curve that are not present in the MFVC. These are either true peaks that are flattened by the smoothing, or outliers arising from errors in computation.

The heist goes bad in 4B, when hostages fight back, a security guard is firebombed and it is evident that Eric the ringleader is insane. Thus we have strong peaks during 4B. During the final conflict in 5A, we have erratic match framing, which shows up as a lower peak. When Zed and Zoe finally ride off into Paris, we have match framing.

In *Killing Zoe*, match framing is used as an audiovisual feature that reinforces the story. The viewers' eyes stay fixed across shot boundaries during scenes where everything is happy or "in control," while the gaze must move quickly to follow the action when things are out of control. Note that there are some violent/intense scenes, such as 4A, which have a low MFVC because there are some violent scenes where

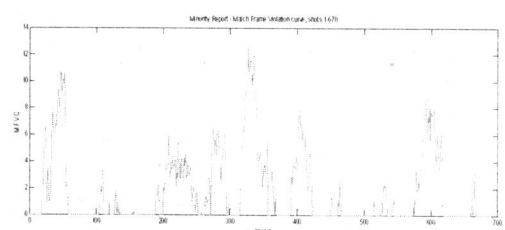

Figure 6: MFVC for the first 670 shots of *Minority Report*.

everything is still under control. Like any filmmaking technique, match framing serves to reinforce the storyline, and its meaning varies from film to film. In *Minority Report*, for example, peaks/troughs in the MFVC may mean something else entirely, Fig. 6.

More results are given in Fig. 7-11 for examples of match frame violation/non-violation.

4. Conclusions and Future Work

There are three limitations that are apparent with this approach to match frame detection. First and most importantly, salience maps as described in [6] do not take motion salience into account. Motion is an important perceptual cue in determining focus of attention. Rapantzikos and Tsapatsoulis [12] are currently attempting to infuse motion into saliency maps. Their efforts, as of yet, have not borne fruit and the authors' own attempts are undergoing preliminary study as part of a larger project. In [12] the authors do however include maps that detect center-surround for human skin. For the purpose of visual attention in film, this is counter-productive as filmmakers put a great deal of effort into ensuring that their intended center of attention is always visually salient in a scene.

The next problem we encounter is in the rigidity of the 3-component GMM, as salient points are not always accurately modeled by a fixed number of components. In future work, we will attempt to model the saliency map as a distribution with a variable number of components using an algorithm such as quasi-GMM [10]. The downside to this approach is that it leads to more complicated distance metrics.

Finally, the limitations are with the different stages of the system itself. False positives in shot detection will not be detrimental to the algorithm, as frames of the same shot will tend to have similar saliency maps, however false negatives can cause false alarms in detection of match frame violations, as the centers of attention often move within a shot. The next stage of the system, the saliency map, is not guaranteed to agree with the viewer's subjective focus. This can again lead to false positives/negatives. Finally, evaluation must be done with different distance metrics, such as the Earth-mover's distance [14].

We believe that with the aforementioned deficiencies rectified, the match frame detection system will make a powerful feature for detection of important moments in motion pictures.

7. References

[1] Bordwell, D. and Thompson, K. *Film Art: An Introduction*. Published by McGraw Hill Companies, 1996. Chapter Eight, "Continuity Editing," pp. 262-78.

[2] Canny, J. *A Computational Approach for Edge Detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence., vol 8. no. 6. pp. 679-698. 1986.

[3] Dayan, P. and Abbot, L.F. *Theoretical Neuroscience* pp. 62-65. The MIT Press, 2001.

[4] Hanjalic, A. *Generic Approach to Highlight Detection in a Sport Video*, IEEE International Conference on Image Processing (ICIP 2003), Special Session on Sports Video Analysis. Barchelona, ES., 2003.

[5] Hanjalic, A. *Multimodal Approach to Measuring Excitement in Video*, IEEE International Conference on Multimedia and Expo, Baltimore USA. 2003.

[6] Itti, L., C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans Patt Anal Mach Intell. 20(11), pp. 1254--9, 1998.

[7] Levaco, R. *Kuleshov on Film*. Los Angeles: University of California Press, p. 8. 1974.

[8] Lienhart, L., S. Pfeiffer and W. Effelsberg. *Video Abstracting*. Communications of the ACM, vol. 40 no. 12, pp. 55-62, December 1997.

[9] Lienhart, R. *Reliable Transition Detection in Videos: A Survey and Practitioner's Guide*. International Journal of Image and Graphics (IJIG), Vol. 1, No. 3, pp. 469-486, 2001.

[10] Lu, L. and H.J. Zhang. *Real-Time Unsupervised* Speaker Change Detection. Proceedings of the 2002 International Conference on Pattern Recognition (ICPR), Quebec City, CA, 2002.

[11] Messaris, P. *Visual Literacy: Image, Mind and Reality*, p. 16. Boulder, CO. Westview Press, 1994

[12] Rapantzikos, K. and Tsapatsoulis, N. On the Implementation of Visual Attention Architectures, http://www.image.ece.ntua.gr/php/pub_details.php?code=228

[13] Rasheed Z. and M. Shah. *Video Categorization using Semantics and Semiotics*, book chapter in *Video Mining*, A. Rosenfeld, D. Doremann and D. Dementhon Eds, Kluwer Academic Publishers, June 2003.

[14] Rubner, Y., C. Tomasi and L.J. Guibas. *A Metric for Distribuations with Applications to Image Databases*. Proceedings of the 1998 IEEE International Conference on Computer Vision, Bombay, India. January 1998, pp. 59-66.

[15] Xie, L., P. Xu, S.F. Chang, A. Divakaran and H. Sun. *Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models*. Pattern Recognition Letters, 25(7):767-775, May 2004

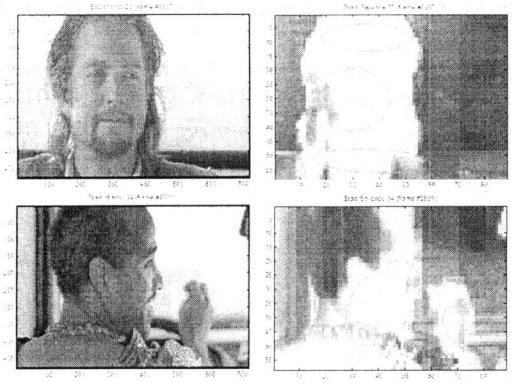

Figure 7: A Match Frame (shots 32-33), where everything is "Supercool"

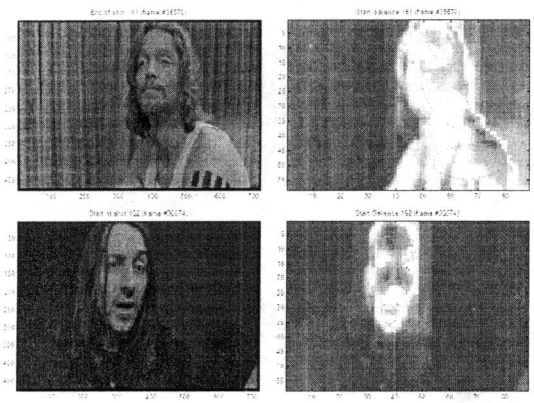

Figure 8: An approximate match (shots 161-162) in a friendly but antagonistic scene

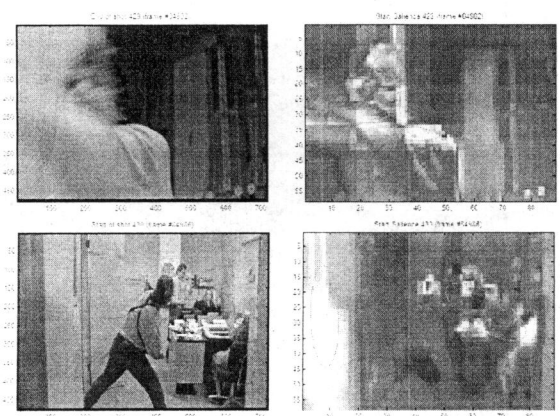

Figure 9: A match violation during the bank robbery

Figure 10: Match framing in *Minority Report*, where Tom Cruise is hunting down a killer

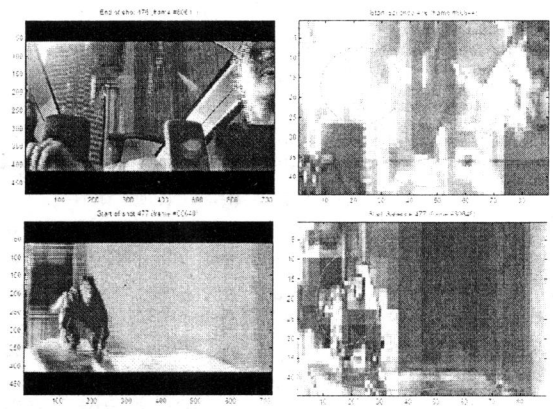

Figure 11: Match frame violation while Tom Cruise is being chased in *Minority Report*