Automatically Selecting Shots for Action Movie Trailers

Alan F. Smeaton, Bart Lehane, Noel E. O'Connor, Conor Brady and Gary Craig
Centre for Digital Video Processing
& Adaptive Information Cluster
Dublin City University
Glasnevin, Dublin 9, Ireland
alan.smeaton@dcu.ie

ABSTRACT

Movie trailers, or previews, are an important method of advertising movies. They are extensively shown before movies in cinemas, as well as on television and increasingly, over the Internet. Making a trailer is a creative process, in which a number of shots from a movie are selected in order to entice a viewer in to paying to see the full movie. Thus, the creation of these trailers is an integral part in the promotion of a movie. Action movies in particular rely on trailers as a form of advertising as it is possible to show short, exciting portions of an action movie, which are likely to appeal to the target audience. This paper presents an approach which automatically selects shots from action movies in order to assist in the creation of trailers. A set of audiovisual features are extracted that aim to model the characteristics of shots typically present in trailers, and a support vector machine is utilised in order to select the relevant shots. The approach taken is not particularly novel but the results show that the process may be used in order to ease the trailer creation process or to facilitate the creation of variable length, or personalised trailers.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Video

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Video analysis, Movie trailers, Machine learning

1. INTRODUCTION

The movie industry is a massive producer of information. According to the Internet Movie Database (IMDB) [12], 10,342 film and video titles were released worldwide in 2001 and according to [22] we have a grand total surviving stock of 328,530 movies representing a total of 740,803 hours of viewing. With such a massive amount of information, ways in which we can preview and rapidly select movies we want to watch have been developed and movie trailers are an example of that.

Movies are an art form rather than a direct recording of something like a sports event, an interview or a TV news broadcast. Most movies feature a trailer or short video advertisement for the movie that is designed to showcase the movie and attract patrons to want to view it by taking excerpts from the movie and re-packaging them into a self-contained short video. In the early days of movie-watching these were shown at the end of a programme in a cinema or theatre as previews of coming attractions and hence were termed "trailers", a term which has stuck. These trailers tend to feature the high points of the movie which are edited together in such a way that they do not give away the storyline or conclusion, and yet act as a teaser to their audience. Trailers themselves can be quite cinematic with their own background music, sophisticated shot transition, and post-produced features such as overlaid text.

Because movie trailers have a creative and artistic aspect, we cannot expect to fully replicate the generation of a movie trailer from a source movie automatically, unless we fully understand the grammar behind what makes both a movie and a trailer, what shots should be included given the movie storyline, and finally how to artistically compose a movie trailer. This is a life skill which we believe is currently not possible to replicate fully automatically. To complicate this even further, movie trailers will vary according to movie genre and the criteria for what makes the set of shots suitable for an action movie trailer would not be the same as the criteria for a trailer for a romance or horror movie trailer, for example.

The hypothesis which we explore in this paper is that whilst we can't create artistic trailers without fully understanding movie grammar, or indeed the movie and trailer grammar for different movie genres, we could help in the process of trailer generation by locating the shots which could be contributors to the trailer of a movie, which would make the task of trailer generation easier and even open up new possibilities. In particular, by concentrating on action movies, where the set of shots which are used to create the trailer could be less difficult to find because they are all-action, we aim to explore how we could automatically locate, from an action movie, selections from the original movie which could go into its trailer. Applications for this could be the generation of variable-length or even personalised movie trailers for the same movie as an alternative

MIR'06, October 26–27, 2006, Santa Barbara, California, USA.

to having to generate a single trailer which is watched by everyone.

The approach we take is fairly standard in that we use low-level video features and we train a standard machine learning algorithm using a support vector machine. The novelty in the work reported is not in the way it is implemented but rather in the development of the application itself. By showing that this is technically feasible thing to do, albeit with a standard approach, we open up new possibilities for applications like this.

The rest of this paper is organised as follows. In the next section we examine movies and their trailers for a particular kind of movie genre, namely action movies, and we introduce the set of movies that we use in our experiments. We then look at related work in video summarisation and trailer generation, which is followed by presentation of the audio and visual analysis techniques used in our work. Section 5 describes our experimental setup and section 6 presents the results of shot selection for action movie trailers. A concluding section completes the paper.

2. ACTION MOVIES AND THEIR TRAILERS

We first introduce the six movies used in our work which can be broadly defined as all of the *action* genre. The movies we used[1] are:

- *Alien* (1979) 116 min., directed by Ridley Scott — a mining ship, investigating a suspected SOS, lands on a distant planet. The crew discovers some strange creatures and investigates.
- *Clockers* (1995) 128 min., directed by Spike Lee — Strike is a young city drug pusher under the tutelage of drug-lord Rodney Little, who likes to hang out with his gangster friends outside the project houses.
- *Crouching Tiger Hidden Dragon* (2000) 115 min., directed by Ang Lee — award-winning foreign language film from China where legendary martial artist Li Mu Bai sets out to track the murderer of his master.
- *Indiana Jones and The Raiders of The Lost Ark* (1981) 115 min., the first of the trilogy directed by Steven Spielberg — archaeologist and adventurer Indiana Jones is hired by the US government to find the Ark of the Covenant, before the Nazis do.
- *Indiana Jones and The Temple of Doom* (1984) 118 min., the second of the trilogy, also directed by Steven Spielberg — after arriving in India, Indiana Jones is asked by a desperate village to find a mystical stone. He agrees, and stumbles upon a secret cult plotting a terrible plan in the catacombs of an ancient palace.
- *Indiana Jones and The Last Crusade* (1989) 127 min., the final instalment of the trilogy, directed by Spielberg — the daring archaeologist and his father search for the Holy Grail and fight the Nazis.

In selecting movies to work with we used movies of the same genre. The Indiana Jones trilogy has three movies which very much stick to the same formula of all-action, although the plot in each is slightly different. However, our work in this area is particularly influenced by the work of Hsuan-Wei Chen and colleagues at the National Taiwan University presented in [10] as their work also used action movies and we shall describe their contribution in the next section. Specifically they used Crouching Tiger Hidden Dragon (2000), Minority Report (2002) and Charlie's Angels II (2003), chosen because they are all action movies but they all have somewhat different primary characteristics with a sci-fi influence in Minority Report, and a more drama-driven influence in Crouching Tiger. Our set of movies were more homogeneous and similar for reasons which will become clear later.

[1]Movie data and descriptions are adapted from the Internet Movie database [12]

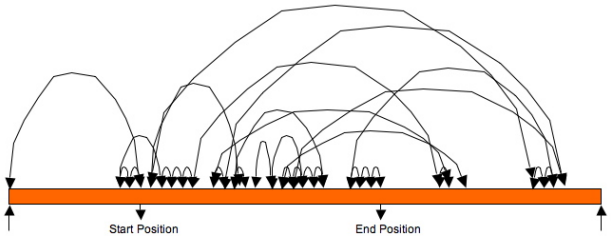

Figure 1: Sequence of shots used in the trailer for the movie *"Indiana Jones, Last Crusade"*, 2 hours and 7 minutes in duration.

For each of the movies we used we located an official trailer, either from the DVD release of the movie, or from an official movie producer's website. These trailers were typically about 2 minutes in duration. We then automatically segmented each of the trailers and the movies into shots, and we then located the shots within the full movie from which the trailer shots were drawn and in this way we established a ground truth for the location of trailer shots for each movie.

It is worth spending some time examining the nature of trailers. Trailers are mostly composed of sub-shots, not the full shots, they use material taken exclusively from within the full movie, and they almost always temporally re-arrange the sub-shots into a different order from the original movie. For example, Figure 1 indicates the ordering of the c.50 shots used in the official trailer for the movie *Indiana Jones, Last Crusade*. We can see from the diagram that the first trailer shot is about one fifth into the movie and the next shots are taken from around that point. At various places in the trailer, shots are taken mostly from the middle and toward the end, occasionally shots are taken from the start of the movie, but the ordering is a complete re-arrangement of the original movie's temporal sequence. The main reason for doing this is to hide the storyline and not reveal the outcome of the plot and we have found this to be a common characteristic among all the action movies we worked with. Figure 2 shows the (unordered) locations of shots appearing in the official trailer for the 6 movies we used and shows that some trailers like *Indiana Jones and the Temple of Doom* use only 25 shots and others like *Crouching Tiger Hidden Dragon* have longer trailers with 70 shots. What is common across most trailers is that trailer shots are taken from throughout the movie's duration except for sections of the graph where there are sharp rises in the plot lines indicating longer sequences of the original move where no shots appear in the trailer.

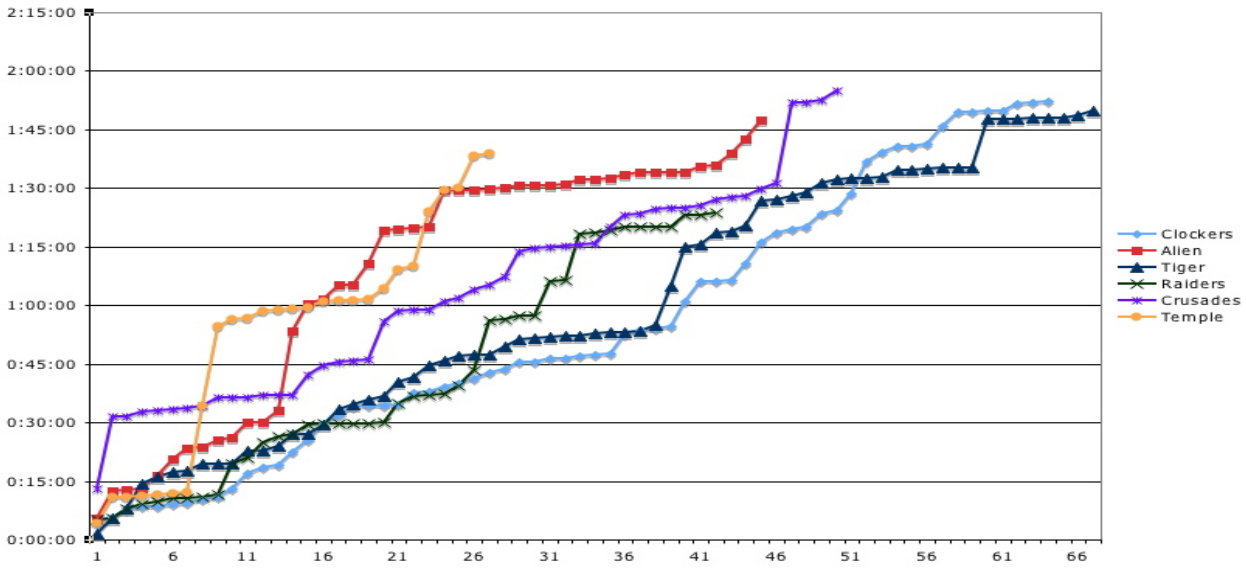

Figure 2: Distribution of Trailer Shots throughout Action Movies

Because of the temporal re-arrangement of shots when generating a trailer, and the fact that it is based on an understanding of the plot and on how to hide the plot using shot re-arrangement, we believe that it is quite difficult to generate a full movie trailer completely automatically, but the isolation of shots which could be used in a movie trailer would be of use to the trailer generation process, and allow trailers of different lengths to be generated more easily.

3. RELATED WORK

The idea of extracting a representative subset from a larger video is not new and has had a number of applications. One of the most popular of these is sports summarisation where many of the analysis techniques focus on the detection of the important events, or highlights, of the particular sports broadcast using audiovisual analysis techniques combined with some form of heuristics or machine learning techniques. For example, Assfalg et al [2] detect free kicks, penalties and corner kicks in soccer matches primarily using camera motion and a hidden Markov model, while [7] look for high motion portions of soccer videos as these typically indicate an important event. [28, 24], analyse the motion information in basketball video in order to determine which team is currently in the attacking part of the court, and use this to detect scores. There have been similar approaches in other sports such as tennis [15], American football [19], motor sports [26], cricket [16] and baseball [9]. One large drawback of many of the sports summarisation techniques presented above is that they are confined to one particular genre of sport. A more efficient summarisation system should be able to detect events over a range of different sports. For example, [27] implements a generic method of detecting highlights in all field sports such as soccer, rugby, Gaelic football and hockey. A number of audio and video features, such as increased audio activity, increased near field visual activity, close up detection, scoreboard activity and crowd detection, were extracted as they constitute the common characteristics that occur in highlights across all field sports. A support vector machine was then used to detect the relevant portions. The main difference between generating sports summaries and movie trailers is that there is no temporal re-ordering of shots in sports summarisation since the aim is detection of highlights rather than some aesthetic product (i.e. a trailer) [33, 32, 34, 35].

There are also many approaches to summarising movies which can broadly be split into *scene-based* summaries and *event-based* summaries. Scene-based summaries (such as [14, 31, 20, 8]) focus on obtaining an index of a movie by splitting it into a set of scenes. The main drawback of these approaches is that little information about the nature of the movie can be garnered from a scene-based structure. Event-based movie summarisation techniques aim to detect shots in the movie that belong to a certain event type. For example, [21] uses colour information in order to detect dialogues in video based on the common shot/reverse shot shooting technique (which results in a sequence of visually repetitive shots). [11] aims to detect both dialogue and action events in a movie, however the same approach is used to detect both types of events, and the type of action events that can be detected is restricted. [23] detects violent events in a movie by searching for visual cues such as flames or blood pixels, or audio cues such as explosions or screaming. Previous work by some of the authors focused on completely summarising a movie by detecting all of the relevant events [18, 17]. Three event classes were chosen (exciting, dialogue and musical) that typically encapsulate all relevant portions of a movie. A range of low-level audiovisual features were extracted and finite state machines were used in order to detect the events.

There are a number of differences between creating a summary of a video, be it a sports video or a movie, and creating a trailer. Firstly, the summaries presented above all tend to focus on extracting full shots from the video, while in trailer generation, sub-shots are used extensively. Also, trailers typically contain some form of temporal re-ordering

of the shots before presentation, while a summary of a video usually follows the same temporal order as the video itself. However, sports highlight detection and event-based movie summarisation techniques are closely related to movie trailer generation as all processes involve selecting relevant shots from the video. Therefore the underlying analysis techniques used by us in our previous movie summarisation work ([18, 17]) can also be utilised in order to detect shots for trailers.

Little work has been reported to date on automatically generating movie trailers from an original movie. The previous work most related to our own, and which has influenced our work is [10], presented at the ACM Multimedia Information Retrieval (MIR) Workshop in 2004. In that work the authors examine the possibility of automatic trailer generation but their analysis of a movie and their composition of a trailer is based on sets of rules or grammars which encapsulate the theory of film composition. They analyse action movies in terms of shot change detection, the MPEG-7 measurement of motion activity intensity and a set of audio features based around audio energy. They automatically extracted and combined these features for an entire movie in order to compute a feature referred to as movie tempo for each shot and then they generated a movie trailer as the concatenation of all the (full) shots from the original movie which have a move tempo value above some threshold.

As we can see from our own analysis above of what constitutes a movie trailer, the work presented in [10] is not what we would conventionally call a movie trailer since there is no selection of sub-shots and no temporal re-arrangement of sub-shots in order to mask or hide the movie storyline. What the authors do to evaluate their generated "sub-movies" is to perform a subjective human evaluation of the"relevance" between human-generated "sub-movies" and their automatic computer-generated movie segments. They also evaluate the "representative" nature of the generated sub-movies in order to indicate how expressive and readable the movie summary is. Their results indicate that further work on trailer composition is needed, according to human judgements, especially across different movie genres.

4. AUDIO-VISUAL ANALYSIS TECHNIQUES USED

In order to perform automatic identification of shots used in a movie trailer, we need to perform audio and/or visual analysis on the movie itself in order to identify features to be used in shot identification. There are a large set of such features we could use and in order to allow some element of comparability with the work of [10] we also use shot length/shot change frequency, we use audio type/class instead of audio envelope/noise and we also use measures to describe the motion within a shot. As trailers for action movies typically contain quite a lot of exciting-type shots, these features were chosen to allow detection of these exciting shots. For example, typically when creating excitement in a movie, a filmmaker will increase the editing pace and amount of motion present in each shot. This may also be accompanied by loud, typically fast-paced, music. This has the effect of startling viewers as a lot of rapidly changing visual information displayed on screen, coupled with fast-paced music, in turn leads to a sense of excitement. We now describe features to detect shots of this nature.

1. We utilise a shot boundary technique in order to generate the basic shot-based structure of a movie. Colour histograms have been demonstrated as a highly accurate and efficient method of comparing images and detecting shot boundaries [6, 3] and are used in our work. A 64-bin Y histogram is extracted from each frame of video and boundaries are detected by locating large inter-histogram differences. This allows us to extract the *shot length* of each shot and although this approach does not work well for gradual shot transitions (such as fades and dissolves) these shot transitions do not tend to occur in action sequences. The rationale for using shot length as a feature is that short shots, in particular sequences of short shots, are likely to indicate action sequences, which is a positive indicator of a likely candidate for a trailer shot.

2. The audio track of a movie is analysed in order to detect the presence of the following categories: *speech*, *music*, *silence*, *speech with background music* and *other audio*. Our rationale for using these audio categories is that music can be indicative of high, or low, points of a movie. As a starting point, four low-level audio features are extracted, namely the High Zero Crossing Rate Ratio, the Silence Ratio, the Short Term Energy, and the Short-Term Energy Variation. The effectiveness of these low-level features in helping to distinguish between speech and music has previously been demonstrated [11, 18]. In order to classify each one second window of audio, a set of support vector machines (SVMs) are used, one for each audio category. The values obtained from processing each one second window are then up-sampled in order to compute the proportion of each audio category for each shot. At the end of this process, for each shot of a movie, there is a value for the percentage of speech, music, silence, quiet music, speech with background music and other audio present (further details of the audio classification process are provided in [18]).

3. For each shot we also detect two motion features, the *motion intensity* and the percentage of *camera movement* present. The motion intensity is an indicator of the amount of motion within each frame of video, and is determined by calculating the standard deviation of the motion vectors. This is firstly calculated for each P-frame of video, and then averaged over the entire shot to give a shot-based value. A method of detecting camera movement is also implemented, which gives the percentage of frames in a shot that contain camera movement (as was previously described in [18]). Although there is some overlap between the motion intensity and the camera movement, both features are required in order to gain a more complete understanding of the type of motion present throughout the movie. The rationale in selecting these features is that shots which have a high degree of motion, especially where that motion is concentrated in a short duration of a long shot, are likely to indicate a high degree of camera and/or object motion, which in turn indicates good candidate shots for inclusion in a trailer.

To summarise, the features used in order to detect shots used in trailers are shot length, motion intensity, and the

amount of camera movement, speech, music, silence, speech with background music and other audio present in each shot. One set of features we did not explicitly use are the visual features taken either from the entire shot, or from the shot keyframe. This includes low-level features such as colour, texture and edges, as well as mid- and high-level features such as occurrence of faces, indoor/outdoor indicators, location indicators (beach, buildings, streets, etc.). We did not use any of these features as they are unlikely to be indicators or highlights for inclusion in a trailer, either alone or in combination. Exceptions to this would be the occurrence of gunshots or other sudden loud noises (detected from audio only), occurrences of explosions (audio and visual detection), or screams/shouts (detected from audio only). This is a topic for future work and for now we want to retain an element of comparability with the work reported in [10].

5. EXPERIMENTAL SETUP

We used the set of 8 features introduced in the previous section, namely shot length, % speech / % music / % silence / % other audio, % speech & background music, motion intensity and the % of shot that has camera motion, and applied these to the original movies digitised into MPEG-1 format

Support vector machines (SVMs) are binary classifiers that will be used in order to determine whether a shot in a movie is potentially part of a trailer or not. It has been demonstrated in other work (such as [27, 25]), that binary classifiers are efficient at selecting relevant shots from video, given a reliable set of input features. SVMs are binary classifiers, the outputs of which are positive values for traile- like shots, and negative for non-trailer shots. The more positive the value for a shot, the more confident the SVM is that particular shot is of use in an action movie trailer. There are many other data classification techniques that have been used in previous work. For example, [17] utilised finite state machines in order to detect events in movies, while [4] used hidden Markov models in order to segment news programs, however both of these forms of analysis focus on detecting continuous parts of the video, rather than isolated shots as is the case when creating a trailer. SVM^{light} was used in this work [13]. This SVM implementation is highly configurable and allows us to choose the optimal parameters (kernel function, cost factor, etc.) for our implementation. A number of configurations were examined, and the one which resulted in best overall performance is presented in this paper.

Evaluation of the performance of our shot selection used the classic measures of precision and recall where a set of shots selected using our trained approach was compared against the ground truth of shots which appear in the official movie trailer. Our approach to using SVM^{light} selects shots in rank order based on their likelihood for inclusion in the original trailer and the specific metric we use for evaluation is $R-Precision$ [1]. Given a ranked list produced as the output of a system to be evaluated, $R-Precision$ is defined as the precision at rank position R, where R is the number of documents or objects relevant to the query. In our case the metric corresponds to a user examining the list of shots offered by our system as candidates for the movie trailer and the user examines this list in order, selecting shots for inclusion until he/she has examined R shots and R is the approximate number of shots the user wants to find, though not all the ones seen at that point will actually be used by the user. For example, in the movie *Indiana Jones and the Last Crusade* the official trailer has 40 individual shots and $R-Precision$ is the precision after the user has looked at the top-ranked 40 shots. This metric has the advantage of being a single-number performance figure though one of the criticisms of it is that it hides the distribution of relevant (and non-relevant) shots throughout the ranking. For example, if system A finds 30 shots from the original trailer as the top suggestions and then suggests another 10 shots not from the trailer then the R-Precision is 0.75. If system B suggests 10 non-trailer shots followed by 30 trailer shots then its R-Precision is also 0.75. Our contention is that the user wants to get (in the case of this example) 40 shots for the trailer and doesn't mind whether the ones to be used appear at the top end or the bottom end of the system's ranking.

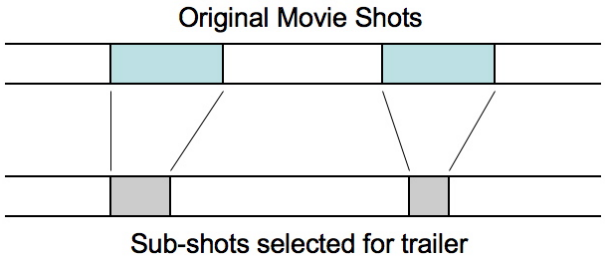

Figure 3: Relationship between movie shots and trailer sub-shots.

When evaluating shot selection we face the issue of how to evaluate sub-shot retrieval, an issue which is illustrated in Figure 3. The issue arises because the movie trailer subsamples the movie shot to select a sub-shot for inclusion so by evaluating in terms of full movie shots retrieved for the user we are masking this. One approach we could take to address this is to evaluate based on the proportion of *frames* from the original movie which appear in the trailer and this would correspond to the way gradual shot transitions are evaluated in TRECVid using $frame-precision$ and $frame-recall$ [30] where the evaluation is in terms of the number of overlapped *frames*. However, as we have already pointed out, we are not trying to select sub-shots from the original movie to appear in trailers as this is an artistic skill. Instead we are trying to identify the full movie shots which could be sub-sampled and edited by an editor into the trailer and thus a more appropriate evaluation measure for us is to use identification of the whole shot even though only a portion of it may be used in the trailer. In our case precision is defined as the proportion of the identified shots for which any sub-shot appears in the official movie trailer and recall as the proportion of the movie trailer shots which are identified by our techniques.

Evaluation of our approach to trailer shot selection was done using a leave-one-out k-fold cross validation. This is a technique used in information retrieval in which a dataset, T, is divided into training $T1$ and testing $T2$ subsets, $T = T1+T2$, training is done on $T1$ and testing on $T2$, and then T is re-divided into different training and testing subsets $T1'$ and $T2'$ and the training and evaluation is repeated, a total of k times [29]. We used our (six) movies in a 6-fold cross validation with each subset of 5 movies being used to train a trailer shot selector and then evaluated on the 6th movie.

Table 1: R-Precision Performance Figures for k-fold cross-validation

Movie	R-Precision
Temple of Doom	0.926
Last Crusade	0.954
Lost Ark	0.971
Crouching Tiger	0.810
Clockers	0.968
Alien	0.897
Average	0.921

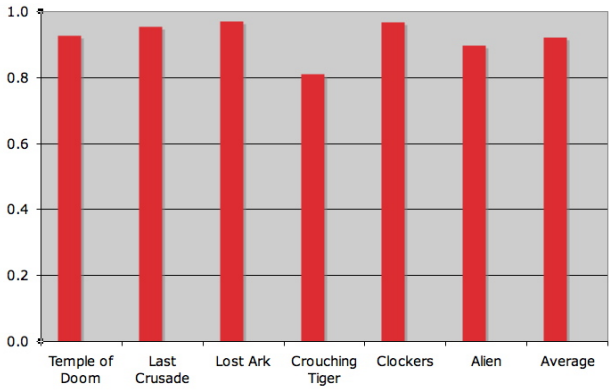

Figure 4: Performance of trailer shot identification in different movies.

Every data point, in our case every movie, gets to be the test set exactly once, and gets to be in a training set $k-1$ times. The variance that we will get among the result sets would be reduced for larger values of k (i.e. more movies) and the disadvantage of this method is that the training algorithm has to be rerun from scratch k times, but for 6 movies that is acceptable.

6. EXPERIMENTAL RESULTS

The results of performing our k-fold cross validation are presented in Table 1 and shown graphically in Figure 4. We used a polynomial kernal function within the SVM and we varied several parameters in the generation of SVMs for shot selection but we do not include those results here, just the results which indicate best performance and are optimised.

The results show several interesting aspects. Firstly, the consistently high results indicate that this approach of selecting shots for action movie trailers is both accurate and reliable. This may be due to the reliance on makers of action movie trailers to adhere to a proven formula of shot selection. As the features detailed in Section 4 are commonly used in order to create excitement, the SVMs could reliably select the correct shots. In general, motion intensity and the amount of music present in a shot were the defining features, although removal of other features also resulted in small decreases in performance.

A second interesting result is that performance for the movie *Crouching Dragon Hidden Tiger* is much worse than for any of the others. This is possibly because that movie is different from the others in that it is not a Hollywood blockbuster all-action movie but is more artistic and although highlights include fight scenes some of those are actually played in slow motion. The second-worst performance is in the movie *Alien* and this can be explained by the fact that the highlights in *Alien* include scary moments where there is tension and build-up and then sudden, unpredictable events as well as human-alien chase and fight scenes. Performance among the other more homogeneous movies is quite good and an overall average $R-Precision$ value of 0.92 is obtained, which is quite high.

One possible danger with our results is that their accuracy could be biased by the use of automatic shot segmentation. A correct classification of a movie trailer shot occurs when the ground-truth trailer sub-shot occurs within the selected movie full-shot. However, the full-shots are based on an automatic shot boundary detection method, not a ground-truth shot segmentation. Hence, if the shot segmentation technique under-segments the movie, the performance may be biased because a larger shot segment may contain multiple scenes and hence be more likely to contain a trailer sub-shot. As a trivial example, if a movie is segmented into just one (very long) shot then all the trailer sub-shots will be sub-shots of this one movie shot, and the accuracy of the result will be 100% precision and recall ! Thus the shot boundary detection method, and its accuracy, is critical to the evaluation approach we have taken. In the experiments we used our own shot boundary detection method which has a precision of more than 90% and a recall of more than 80% when tested on a very heterogeneous collection [5] so we believe the situation of under-segmentation does not arise.

7. CONCLUSION AND FUTURE WORK

This paper presented an approach for automatically selecting shots from action movies in order to assist in the creation of trailers. The implementation approach used – analysis using low-level image and video features used to train a support vector machine – is fairly commonplace but the novelty in the work is in the application itself, not in the way it is implemented. At the beginning of the paper we hypothesised that the process of shot selection for action movie trailers could be automated and, as the results of the experiments in section 6 show, this has been validated. Consistently high precision and recall values across all movies indicate that there is a reliance on filmmakers to use particular types of shots in trailer generation, and that our approach may assist in this shot selection process.

The next step for this work is to address personalising the trailer for individuals so allow generation of longer, or shorter, trailers depending on a person's preference, and to apply this technique to the generation of trailers for genres other than action movies. In order to do that we may need to expand the collection of movie features we extract automatically. For example, the detection of gunshots, explosions, screams, etc. may assist in the shot selection process for action and/or horror movies, and by analysing the music in a movie and locating the areas where emotional music is played, it may be possible to locate the shots to be used when creating trailers for dramas or romantic movies.

One drawback to our approach that is not addressed here, but is an area for future work, is the selection of sub-shots rather than whole shots. As previously mentioned, many of the shots within trailers are in fact sub-shots from the movie.

One possible method of selecting these sub-shots may be to examine the motion present in a shot and remove parts of the shot with low amounts of motion.

Acknowledgements

This work is partly supported by Science Foundation Ireland under grant number 03/IN.3/I361. We are grateful for insightful comments and feedback provided by reviewers.

8. REFERENCES

[1] J. A. Aslam and E. Yilmaz. A geometric interpretation and analysis of R-Precision. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 664–671, New York, NY, USA, 2005. ACM Press.

[2] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala. Soccer highlights detection and recognition using HMM's. In *Proc. IEEE ICME 2002*, 2002.

[3] J. S. Boreczky and L. A. Rowe. Storage and retrieval for image and video databases (SPIE). In *Comparison of Video Shots Boundary Detection Techniques*, 1996.

[4] S. Boykin and A. Merlino. Communications of the acm. In *Machine Learning of Event Segmentation for News on Demand*, February, 2000.

[5] P. Browne, A. F. Smeaton, N. Murphy, N. O'Connor, S. Marlow, and C. Berrut. Evaluating and combining digital video shot boundary detection algorithms. In *IMVIP 2000 - Irish Machine Vision and Image Processing Conference*, 2000.

[6] P. Browne, A. F. Smeaton, N. Murphy, N. E. O'Connor, S. Marlow, and C. Berrut. Evaluating and combining digital video shot boundary detection algorithms. In *Irish Machine Vision and Image Processing Conference*, 2002.

[7] R. Cabasson and A. Divakaran. Automatic extraction of soccer video highlights using a combination of motion and audio features. In *Symp. Electronic Imaging: Science and Technology: Storage and Retrieval for Media Databases*, volume 5021, pages 272–276, Jan 2002.

[8] Y. Cao, W. Tavanapong, K. Kim, and J. Oh. Audio-assisted scene segmentation for story browsing. In *Proceedings of the International Conference on Image and Video Retrieval*, 2003.

[9] P. Chang, M. Han, and Y. Gong. Extract highlights from baseball game video with hidden Markov models. In *Proc. IEEE Int. Conf. on Image Proc. (ICIP)*, 2002.

[10] H.-W. Chen, J.-H. Kuo, W.-T. Chu, and J.-L. Wu. Action movies segmentation and summarization based on tempo analysis. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 251–258, New York, NY, USA, 2004. ACM Press.

[11] L. Chen, S. J. Rizvi, and M. Ötzu. Incorporating audio cues into dialog and action scene detection. In *Proceedings of SPIE Conference on Storage and Retrieval for Media Databases*, pages 252–264, 2003.

[12] The Internet Movie Database - IMDb. http://www.imdb.com/, Last visited June 2006.

[13] T. Joachims. *Making Large-Scale SVM Learning Practical*. MIT-Press, 1999.

[14] J. R. Kender and B.-L. Yeo. Video scene segmentation via continuous video coherence. In *Proceedings CVPR*, pages 167–393, 1998.

[15] E. Kijak, L. Oisel, and P. Gros. Temporal structure analysis of broadcast tennis video using hidden markov models. In *Electronic Imaging: Science and Technology: Storage and Retrieval for Media Databases*, volume 5021, pages 277–288, 2003.

[16] M. Lazarescu, S. Venkatesh, and G. West. On the automatic indexing of cricket using camera motion parameters. In *Proc IEEE ICME 2002*, 2002.

[17] B. Lehane and N. E. O'Connor. Workshop on image analysis for multimedia interactive services (WIAMIS), Incheon, Korea. In *Movie Indexing via Event Detection*, 2006.

[18] B. Lehane, N. E. O'Connor, and N. Murphy. Dialogue scene detection in movies. In *International Conference on Image and Video Retrieval (CIVR), Singapore, 20-22 July 2005*, pages 286–296, 2005.

[19] Li and M. Sezan. Event detection and summarization in American Football broadcast video. In *ESymp. Electronic Imaging: Science and Technology: Storage and Retrieval for Media Databases*, volume 4676, pages 202–213, 2002.

[20] Y. Li and C.-C. J. Kou. *Video Content Analysis using Multimodal Information*. Kluwer Academic Publishers, 2003.

[21] R. Lienhart, S. Pfeiffer, and W. Effelsberg. Scene determination based on video and audio features. In *In proceedings of IEEE Conference on Multimedia Computing and Systems*, pages 685–690, 1999.

[22] P. Lyman and H. R. Varian. How much information ? http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/, Last visited June 2006.

[23] J. Nam, M. Alghoniemy, and A. H. Tewfik. Audio-visual content-based violent scene characterization. In *Proceedings of International Conference on Image Processing (ICIP)*, volume 1, pages 351–357, 1998.

[24] S. Nepal, U. Srinivasan, and G. Reynolds. Automatic detection of goal segments in basketball videos. In *Proc. of ACM Multimedia*, pages 261–269, 2001.

[25] N. O'Hare, A. F. Smeaton, C. Czirjek, N. E. O'Connor, and N. Murphy. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2004, Montreal, Quebec. In *A Generic News Story Segmentation System and its Evaluation*, 2004.

[26] M. Petkovic, V. Mihajlovic, M. Jonker, and S. Djordjevic-Kajan. Multi-modal extraction of highlights from TV Formula 1 programs. In *Proc. IEEE ICME*, 2002.

[27] D. Sadlier and N. O'Connor. Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1225–1233, 2005.

[28] D. Saur, Y.-P. Tan, S. Kulkarni, and P. Ramadge. Automated analysis and annotation of basketball video. In *Symp. Electronic Imaging: Science and*

Technology: Storage and Retrieval for Image and Video Databases, volume 3022, pages 176–187, Jan 1997.

[29] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[30] A. F. Smeaton. Large Scale Evaluations of Multimedia Information Retrieval: The TRECVid Experience. In W.-K. L. et al., editor, *CIVR 2005 - International Conference on Image and Video Retrieval*, volume LNCS 3569, pages 11–17, Singapore, July 2005. Springer.

[31] H. Sundaram and S.-F. Chan. Determining computable scenes in films and their structures using audio-visual memory models. In *ACM Multimedia 2000*, 2000.

[32] D. Tjondronegoro, Y.-P. P. Chen, and B. Pham. Sports video summarization using highlights and play-breaks. In *MIR '03: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 201–208, New York, NY, USA, 2003. ACM Press.

[33] D. Tjondronegoro, Y.-P. P. Chen, and B. Pham. The power of play-break for automatic detection and browsing of self-consumable sport video highlights. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 267–274, New York, NY, USA, 2004. ACM Press.

[34] X. Tong, Q. Liu, Y. Zhang, and H. Lu. Highlight ranking for sports video browsing. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 519–522, New York, NY, USA, 2005. ACM Press.

[35] H. Xu and T.-S. Chua. The fusion of audio-visual features and external knowledge for event detection in team sports video. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 127–134, New York, NY, USA, 2004. ACM Press.