The computer can tell what’s exciting

Choosing videos: emotion curve replaces storyline

Editors always have to dig into the image archives to put together documentaries, or to illustrate television news stories. Television companies have masses of video material at their disposal. The archives keep on growing so that locating the right bit of footage on people who are not in the news daily is becoming more and more difficult.

Christian Jongeneel

As many viewers will know from experience, the thrilling description on a video cover offers no guarantee of suspense in the video itself. The exciting bits may add up to only a few minutes of the 90 minutes or so running time, and that is something you would have liked to have known in advance. Dr. Alan Hanjalic at the Information Technology and Systems (its) faculty of Delft University of Technology has collaborated with Dr. Li-Qun Xu of British Telecom to develop a method that will let a computer analyse the video content for suspense, melancholy, mirth, or any mix of these emotions. On top of this the Delft researcher has been working on a system that will automate access to video archives. Over the years, film and television companies have built up enormous archives. The Dutch Audiovisual Archive alone contains some 900,000 hours of analogue audiovisual material. The only way to make this material accessible is to look at the video frames, make a record of what you see, and store the record in a computer. Hanjalic’s system can classify video stills as well as moving video frames. It can pick...
out the goals in a soccer match with no trouble at all. News broadcasts are automatically cut into sections and arranged according to subject.

Computers know how to handle hard data. They can spot the colours in video frames and detect object and camera movement, or the frequencies of the soundtrack. But they have no idea at all of what is being shown, so they cannot tell the difference between a peaceful landscape with a winding stream and twittering birds and an action-packed war movie full of explosions with bullets flying all over the place. To a computer it’s all the same. At least, until recently. After five years, Hanjalic has the workings of a system that can also look at the content of video frames. ‘There is a difference between the measurable properties of audiovisual material and what it means’, he explains, ‘and this is called the semantic gap. This gap will never be bridged, but you can make it smaller. The idea is to reduce the search space.’ To give a more concrete example, Hanjalic once saw some broadcasting archivists working for a Portuguese television company. These people had to wade through scores and scores of tapes to make a record of what could be seen and heard. Work of this kind could be done much more efficiently if a computer could provide suggestions. Even if the computer were to be right in only half the cases, the amount of work would be reduced significantly. And there are more applications, the best-known of which is the anti-violence microchip for television sets which can indicate whether a broadcast is too violent to show to children. A device like that (which does not exist yet) must also be capable of assessing the content of a video. ‘When you begin your analysis, you first have to ask what exactly you want to extract from the images’, Hanjalic explains. ‘Are you trying to recognise the objects and scenes captured by the camera – is it an animal or a landscape – or do you want to know what the viewer’s emotions will be after watching the video? The latter is what I’ve been working on most recently. Dr. Xu and I are the first to do so in Europe.’

**Key frames**

Obviously it is virtually impossible not to mention undesirable to scan a video frame by frame to see whether it qualifies as suspense, comedy, or what have you. Video material has to be judged by looking at an entire scene. To do that you need to know where each scene starts and ends. Scenes consist of a series of shots (frames shot by a single camera), and even detecting the start and end of a shot is far from easy. ‘As long as the camera and the objects in the picture move at normal speed, everything is pretty straightforward’, Hanjalic explains. ‘You compare a succession of video frames, and if they look roughly the same, they’re all part of the same shot. People have been working on this principle for a decade, but there still isn’t an algorithm that will work in every case. Sometimes there is simply too much difference between consecutive frames, for example if flashes of light occur, or abrupt movements close to the camera. What’s more the transitions between shots aren’t always abrupt. If the shots fade into each other,
the computer will miss the shot change. If all fadeovers were of the same general form and length, the computer might be able to manage, but that is not the case. In other words, there are no rules for the length and form of scenes and shots.

Hanjalic introduced a number of improvements in the existing methods for distinguishing different shots using what is known as statistical detection theory. This allows the decision about a certain hypothesis («shot change» or «no shot change») to be taken at a certain time on the strength of a combination of different statistical parameters. The first parameter, the shot length, keeps a record of the number of frames that have passed since the last recorded shot change. By weighing this parameter against previous knowledge about the conventional length of shots, the results are improved. Ultra-short shots are rare. If the computer keeps finding one-second shots, there is probably something else going on, like a scene in a disco with a strobe light. The second and third parameters use the differences between consecutive frames. The second parameter tests only to see if the measured value is sufficiently high or low to support the first or second hypothesis. In simple terms, it makes use of the fact that, if one frame is very different from the previous one, we probably have a shot change. The third parameter compares the pattern created by a number of successive readings with the expected shot-change pattern predicted by a model. If the patterns match, we probably have a shot change. The system then selects a number of key frames from each shot. These frames are representative for the shot as a whole. In a steady shot, a single key frame may suffice, but in an action shot several may be needed. Again the selection takes place automatically, based on the amount of movement between consecutive frames.

Four weddings

The next step is to group the shots into a scene. ‘A scene is characterised by a certain location and certain characters’, Hanjalic says. ‘This means that you have a certain colour composition, of which you can assume that it remains consistent over several shots. A good example is a conversation between a number of persons, in which the camera switches from person to person. You start with the first shot featuring person A. If the key frames in that shot match the key frames in the sixth shot for fifty to seventy percent – which is probably another shot featuring A – the odds are that you have a single scene.’ Assuming for the moment that the director did not edit the shots together in random fashion, the five intervening shots form part of the same scene. You then scan the rest of the video for shots with matching key frames, and you may find a couple of dozen shots that form a unit, i.e. there are no more intervening loose shots, and after that there are no more shots similar to the shots in this unit. You have just found the start and end of the scene. ‘We have tested this model on two films, Jurassic Park and Four Weddings and a Funeral’, Hanjalic continues. ‘In sixty to seventy percent of all cases, the computer managed to find the scene limits. The remainder of the scene transitions did not fit the model I was using, and so the computer failed.

Schematic storyline of a documentary or film.

Sequence of frames from the film, Four Weddings & A Funeral, in which a camera change takes place. Each series of frames before and after the change is called a shot. To facilitate the comparison of different shots, a number of key frames are extracted. These frames represent the visual contents of the shot.
to spot them.

**Newsreels**

So far, Hanjalic’s research had followed the beaten track for image recognition, but his work actually serves a different purpose, which is to classify content. Before he went in for emotion recognition, Hanjalic sat and watched news broadcasts. He wanted to find a way to automatically edit these into separate items. The techniques outlined above simply aren’t up to this task, since the recurrent newreader shots would trick the system into classifying all the shots as part of a single scene. ‘We started by looking at the video signal,’ Hanjalic explains the approach taken by his team. ‘The newreader is easy to pick out. When he appears, that is a strong clue for the end of a report. On the other hand, the newreader does not appear between short news items, so images alone are not good enough. This is why we also looked at the sound pattern. A pause of a few seconds is another indicator of a possible switch from one item to another.’ Once the potential subject boundaries have been spotted, the next task is to determine the subject matter of each segment between two potential boundaries. This still requires human intervention. ‘What you need is a database of subjects,’ Hanjalic says, ‘and to build this, you take a number of articles about a certain subject. A mechanism then decides how specific to the subject certain words are. Each word is then assigned a weighting factor.’ The sound of each news item is then converted into text using speech recognition software. The frequency with which certain words appear is compared with the distribution of words in the texts held in the database. If the word distribution in the segment matches that of subject X in the database, the segment probably is also about X. ‘Modern news footage often makes the task even simpler, because news stations send the text along with the images,’ Hanjalic says, ‘so you can skip the speech recognition stage. Anyway, future broadcasts will probably include subject indicators. My system is intended for use in archives.’ Hanjalic has tested his system on a number of different news broadcasts, and his system scored very well. ‘If the subject has been sufficiently trained, i.e. if the texts in the database are representative, the classification system works well.’

**Heartbeat**

Hanjalic followed a similar path to find a way of recognising video segments based on the emotions they evoke. The main difference is that the boundary between two news items is more or less a given, whereas the distinction between emotions is much more difficult to pinpoint. Therefore, choices had to be made.

He used an existing model which divides human emotions along three axes: control, arousal, and valence. Models like this are commonly used for psychological research into the relationship between emotional observations and their physiological effects on the body. Control over the situation plays a major role in real life, but not when watching a video, so that aspect could be left out. The result was a two-dimensional model covering arousal (from calm to excited) and valence.
This model is popular among psychologists, because both factors can be measured. Arousal can be measured, for instance, through skin resistance, while the degree of pleasure is linked to, amongst others, the rhythm of the heart and movements of facial muscles. Television programme makers also rate the emotion factor high, for it is a well-known fact that excitement is the primary factor determining the appreciation of sporting events. A system that could automatically select highlights from a soccer match would be very popular with the television industry. If you could make an emotion diagram of a video, it would be much easier to cater for individual tastes. All in all, it is a subject that is sure to catch the interest of the entertainment industry, which is why Hanjalic carried out part of his work in this field at the British Telecom research laboratory. BT is looking for ways to personalise the delivery of television programmes at home.

**Saving Private Ryan**

The question that still needed to be answered was how to quantify arousal and valence from a video. ‘For arousal, I found three characteristics,’ Hanjalic says, ‘the first of which is movement. More movement usually means that exciting things are going on. The same goes for the length of the shots. A long shot points to a calm scene. The third characteristic is sound energy at higher frequencies.’ The latter boils down to people screaming. What Hanjalic is doing in fact, is to detect the tricks that film directors use to build up suspense. The role of sound is easily demonstrated by the noise from a soccer match: everybody knows that sports commentators talk with more excitement in their voices when a goal is being scored. Radio commentators even exaggerate this audio aspect in order to compensate for the lack of visual information in bringing excitement to the listeners. Hanjalic used his method on one of the soccer matches between Spain and Sweden, and on a film, Saving Private Ryan. The results closely matched the anticipated experiences of a human viewer. During the football game, goals and opportunities to score resulted in obvious peaks in the diagram. For valence, Hanjalic has so far managed to find one feature, which is the fundamental frequency (tone) of speech signals. Low voices indicate a depressed atmosphere, high voices point to optimism. ‘I have tested this on some footage from Saving Private Ryan, and it turned out to work well. You can see that the curve is negative all along the way. There is only one point where the system says that the film is pleasant, and that is the scene where the main characters are walking along, cracking jokes.’ Hanjalic has filed a patent on the system together with his colleague, Dr. Li-Qun Xu of British Telecom. He shows a diagram in which arousal and valence have both been plotted. The course of the film becomes clear immediately with a negative and calm start (for those who have seen the film: the part analysed skips the horrific landing scene), then rises slightly in arousal and valence, shoots back to negative though more exciting, reaches a hair-raising climax, and then gradually sinks once the first frame showing the newsreader has been located, it can be used to track down all subsequent frames containing the newsreader. The plot shows the quality of the match between the first newsreader and all other frames of the news programme. Abrupt peaks indicate a close match and almost certainly indicate a shot of a newsreader. The second curve T(k) represents the threshold value that is automatically adapted to the curve s(k), acting as a filter for high peaks.

Graphical representation of a speech signal from a news broadcast. The moments of silence are fairly easy to detect. Since periods of silence often indicate a change of story, they are used, like shots of anchormen, to mark potential boundaries.

Now that a number of potential story changes have been spotted, the next task is to determine the actual changes, using written text from the news stories. First of all, each sequence of images between two possible changes is isolated. Next, a list of potential subjects is added to the sequence using the text as a basis. The probability of a subject actually forming part of the series of images is determined by the number of words spent on the subject in the accompanying text, and the relevance of those words to the subject. The relevance is determined by an expert system with numbers of texts from old news footage and Internet news pages.
back to calm and reasonably positive.

**Cannibal**

Hanjalic is the first to admit that his system is far from perfect. Films that do not follow the accepted rules are hard to classify. A cannibal calmly eating his way through the contents of a human head, as shown recently in Hannibal, will slip past the arousal detection. There is still a long way to go before the system is practical. ‘I’m convinced that this is one of the research directions of the future,’ Hanjalic says. ‘One of the most important things to be done is to develop more features to improve the classification of emotions. There will also have to be tests with users in order to determine how well the opinion of the system matches the impressions of the viewers.’

For more information please contact Dr. Alan Hanjalic, phone +31 15 278 3084, fax +31 15 278 1843, e-mail : [http://www.delfoutlook.tudelft.nl/info/mailto_a.hanjalic@its.tudelft.html](http://www.delfoutlook.tudelft.nl/info/mailto_a.hanjalic@its.tudelft.html), or Prof. Dr. Ir. Inald Lagendijk, phone +31 15 278 3731, e-mail : [http://www.delfoutlook.tudelft.nl/info/mailto_r.l.lagendijk@its.tudelft.html](http://www.delfoutlook.tudelft.nl/info/mailto_r.l.lagendijk@its.tudelft.html)
Parents (and others, no doubt) wish to prevent their children from being confronted with violent and otherwise shocking scenes in films on television and on video. The development of an anti-violence microchip has been the subject of debate for a number of years now. It has not been developed yet, because there is no algorithm that can detect violence, and because the broadcasting companies are still failing to include an electronic stop sign in the television signal that only parents can override using a special code. Hanjalic has developed a method that can say something about a video’s content by using an algorithm that can analyse the emotional properties of the video material.

At the University of Amsterdam, a sensor suit is used to measure the effect of visual stimuli on test subjects.

For some time psychologists have been using a model to represent the human emotions along three axes: control, arousal, and valence. Research results show that all our emotions fall within the grey area. In other words, no stimuli can induce an emotional condition outside the grey area.
Emotions evoked by radio, television, computers, and sound tend to be passive, because the control component is very slight. Therefore, Hanjalic’s algorithm measures arousal and valence only. If all the readings are plotted without the control component, they fall within this curve.

An historic picture showing the famous Dutch soccer commentator Herman Kuiphof at a match in 1969. The archives of Dutch television companies alone contain hundreds of matches that are waiting to be digitised. As in future matches, viewers will want to be shown automatically selected highlights from the game. Of course, advertisers will be less pleased. One of the indicators of a climax is the way the reporter raises his voice.
Valence curve of a fragment from the film, Saving Private Ryan. This part of the film is generally regarded as fairly downcast, with the exception of the halfway part, where positive valence readings occur.

Arousal curve of a fragment of the Spain vs. Sweden football match, measured using Hanjalic’s algorithm.

Arousal curve of a fragment from Saving Private Ryan. The two climaxes in this fragment correspond with combat situations.
Emotion curve of the above fragment from Saving Private Ryan, according to Hanjalic’s algorithm. From the trace of the curve in the two-dimensional space, we can see that fragment A is rather exciting and grim, B is extremely exciting, C is extremely bleak, D causes mixed feelings with little excitement, and E is cheerful.