1 Introduction

The research into image databases and image indexing and retrieval has led to the creation of a number of useful tools for similarity retrieval for images [5, 9, 10]. Application of these tools to video is possible, but the principles embodied in the tools do not yield a useful query system. Previous work on video indexing and retrieval [21, 24, 16, 25, 11, 7, 22, 4] has most commonly relied largely on one aspect of video, be it vision or sound, and has been restricted to low-level processing. The results of this processing can then be used for classification, with the goal of detecting some form of structure within video. This allows summary of video, thus permitting a restriction of the segments needed for browsing. Other work uses closed captions from news programs to perform natural language understanding for semantic annotation [3, 12]. The used of closed captions limits the applicability of this work, as few programs are produced with this additional information.

In this paper we describe a collection of tools and their application to detection of structure in a news broadcast. In particular, these tools are used to break the broadcast into segments, each of which corresponds to a single topic of discussion. These segments are classified further by labelling each individual shot as one of

- anchor person or reporter,
- footage with a voice over,
- sound bite,

which gives a clear indication of structure within the video.

This work differs from earlier work in that it employs not only low-level processing, but combines results from various modes of processing, along with initial deductions about structure within video, to apply higher level processing in a directed manner. This allows an iterative approach to be used, with alternating processing and deduction employing progressively more complex computation as the interpretations become more finely focused. Domain knowledge allows the direction of processing onto portions of the data set most likely to provide rewarding results. This approach makes good use of resources by using domain knowledge and simple initial processing to carefully chose video segments for more detailed processing. The aim of this work is to allow automated annotation of video, which will allow intelligent construction of summaries for large video databases. The particular target area is news broadcast and footage, such as that kept by major news companies. The annotation created will break the video into segments of homogeneous topic, and further label shots as anchor or footage. A typical summary might then be created by providing a thumbnail of each anchor person or reporter present in a section of video. Given the large volume of video data retained for such applications, and the volume captured at each moment, this could result in a large reduction in unproductive human time.

A similar piece of work has been attempted by Huang *et al* [8], however, this work assumes that the set of speakers is known and a suitable training corpus of vocal samples exists. Use is also made of closed caption text for semantic unit extraction. These specialisations, in addition to the strict assumptions made as to the shot syntax of the individual storys makes this work inapplicable to the majority of video streams.

2 Visual processing

There are a number of tools that are commonly employed in the analysis of video streams. These tools are most often employed to assess similarity of images from the video stream in a suitable attribute space. The similarity measures within the video stream can then be used to detect shots that are related by image similarity [26, 24]. In this paper we present a novel approach that uses similarity within the video stream, along with domain knowledge of *shot syntax*, to deduce narrative structure within a news video stream.

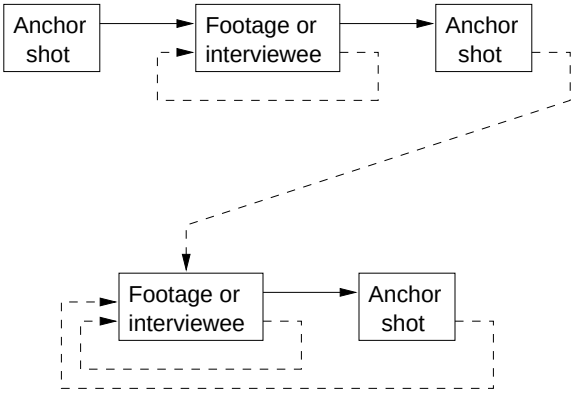

Figure 1: Shot syntax of interviews

The term shot syntax was coined by Bolle et al [2] and describes the regular structure of camera parameters employed to capture a particular type of semantic content. Perhaps the clearest example of a regular shot syntax is seen in interviews. Within an interview segment it is generally the case that the initial shot will be an introduction by the interviewer. This shot will be followed by either a shot of the interviewer and the interviewee, or a shot of the interviewee alone. Subsequent shots will be of either;

- the interviewer,
- the interviewee,
- a mid–range shot of the two people involved,
- background footage.

The structure tends to alternate between one shot of the interviewer, and one or more shots of another type. This structure is indicated in the diagram of Figure 1, where the solid lines indicate required components, and the broken lines represent optional paths through the syntax. This repetitive structure is generally adopted for interviews as it has been proven the most effective method of producing this type of video content.

The presence of such repeated structure within a video stream makes detection of repetition in shot settings a useful step in the grouping of shots into meaningful segments. The regular structure displayed in Figure 1 makes it useful to search for repetitions of anchor or reporter segments. The term anchor shot will be used here to refer to a shot of a single person that is repeated at both ends of a segment and within a segment. This may be a shot of an actual anchor person, or a shot of a reporter who is the presenter for a particular story.

The search for anchor shots within a video stream is facilitated by the fact that the purpose of such shots is to provide continuity to a narrative. In order to achieve this, anchor shots are usually captured with as consistent a background and camera parameters as possible. This presents a linking shot to the viewer before crossing to a separate location, footage or an interview shot. Thus the purpose for which anchor shots are intended makes the task of detecting them simpler as it ensures a consistent visual setting.

2.1 Detection of visual similarity

The initial approach for detection of anchor shots used colour coherence vectors (CCVs) [14] to detect similarity between frames sampled from a video stream. This similarity measure was then used to group sampled frames into repeated shots and other shots. A set of shots which exhibit visual similarity

(a) Frame 111 (b) Frame 112 (c) Frame 113 (d) Frame 114

(e) Frame 115

Figure 2: Poor CCV performance

	111	112	113	114
112	5886			
113	25759	25559		
114	7839	4681	25544	
115	4326	7570	25721	8303

Table 1: Similarity table for frames 111–115 using CCV

	111	112	113	114
112	71112			
113	71410	5374		
114	70844	8220	5454	
115	71430	16644	14718	15260

Table 2: Similarity table for frames 111–115 using spatial histogram

are said to belong to the same visual similarity group. However, when CCVs where applied to this task it was found that there are a number of scenarios that occur frequently in news and interview video for which CCVs do not perform well.

The prime example of this is shots of a single face, usually an interviewee, in which the face rotates slightly. In this case a slight rotation of the face can cause a large change in colour coherence properties. The video frames in Figure 2 and the similarity values in Table 1 illustrate this problem. Figure 2 presents five frames from a section of a CNN news program, depicting a wide shot of the studio and anchors (Figure 2(a), frame 111), and four frames of a single person talking (Figures 2(b), 2(c), 2(d) and 2(e), frames 112 to 115). Table 1 shows the similarity values computed for frames 111 to 115 using CCVs. The best threshold determined for this video is 10000, which performs well for the majority of the video. When this threshold is applied in this case frames 111, 113, 114 and 115 are grouped together, with frame 112 classed as different. The desired results is for frames 112 to 115 to be grouped as similar with 111 placed in a separate group. In fact the similarity values returned for frame 112 compared to frames 113–115 are extremely high, even thought the change is perceptually small. The similarity value for frame 111 compared to frames 113–115 is also significantly less than a reasonable threshold value for the majority of the video.

Due to this difficulty a different measure for similarity was employed, in which each frame is broken into 12 subframes and a colour histogram is computed for each. Each frame is partitioned into four parts along the horizontal axis and three along the vertical axis. For two frames f_i and f_j, the histogram difference $\delta_{x,y}$ is calculated between each pair of corresponding subframes. The sum of the subframe differences is used as the final distance measure Δ_{ij} between frames f_i and f_j. This is expressed in equation 1.

$$\Delta_{ij} = \sum_{x=1}^{4} \sum_{y=1}^{3} \delta_{x,y} \tag{1}$$

Table 2 shows the similarity values computed for the frames in Figure 2. In this table there is a clear separation between the similarity values for frame 111 to 112–115, and the values for pairs of frames from 112–115. The threshold used for this measure is 30000, which shows a significant separation from both the sets of similarity values. The separation between similar and dissimilar frames is considerably more pronounced for the histogram measure than for the CCV measure. The similarity values for frames 112–115 reflect the visible differences, such as frame 115 having a larger similarity value due to the introduction of a caption in the lower part of the frame. The perturbation to the similarity measure due to such factors is however much less than that from the different shot.

Modification of the CCV algorithm to use a similar partitioned approach was attempted, but proves to be unworkable. The CCV algorithm uses a threshold for object size to determine coherence. If the image is partitioned, then one of two approaches must be used, either

1. the threshold is reduced, or

2. retain the threshold.

There are problems with both possibilities. If the threshold is reduced then the measure is changed dramatically. Reduction in the size of object which is required for a coherent object alters the similarity measure greatly, classifying images with much smaller regions of colour as having a high degree of coherence. This reduces the discrimination possible with the CCV algorithm. If the threshold is not altered, then objects which appear across the boundary of a partition will be divided and may be classified as incoherent. Once again this alters the measure in a way that greatly reduces its utility. In both cases the measure is no longer consistent, and can become ill conditioned with small changes to an image, such as an object crossing a partition boundary. The experiments with these two modified CCV approaches showed them to be poor measures of image similarity.

2.2 Separation of anchor and repeated shots

There are other shots besides anchor shots which will be repeated during a broadcast, such as the logo of the news station, adverts which are repeated and footage used as a preview for stories in later programs. In order to aid in distinguishing between anchor shots and other repeated shots, face detection is performed on all repeated shots. This is the initial example of directed application of higher level processing. Face detection is only reliable in constrained applications, so a number of properties of anchor shots are used to simplify the problem. The following properties exhibited by anchor shots make face detection more reliable:

- the face is turned directly towards the camera,
- the face dominates the shot.

Face detection applied to differentiate between anchor shots and other repeated shots can therefore be restricted to searching for large faces. The majority of artifacts the are often detected as faces are small relative to the frame size and may be eliminated on size alone. Limiting the search to only those faces which are directly addressing the camera further reduces the error rate.

Shots that display repetition properties which match the shot syntax of an anchor shot, and have a consistently visible face are determined to be anchor shots. Assuming temporal consistency we can further reduce the error rate from face recognition by discarding faces that move rapidly or erratically. This is particularly useful for discarding footage of a person delivering a speech, such as politicians or other public figures, and also for discarding advertising. Temporal consistency is also applied to the colour histogram measure by using the average histogram for each similarity group to represent its attribute set. This limits the spread of a single group by preventing a sequence of frames, each of which has a small histogram difference from the previous frame, remaining part of the group even though the cumulative histogram difference diverges from existing members of the similarity group.

These two initial processing steps of shot similarity detection and face detections provide information which is combined with knowledge of shot syntax to make initial deductions about the narrative structure of a video. Each shot is assigned a preliminary label, either as an anchor shot, or a non–anchor shot.

3 Aural processing

One approach to further refining the annotation of shots within a video stream, is to examine the audio stream associated with the video. The major impediment to processing sound from news video is the difficulty of ensuring clean audio samples. Modern voice recognition systems produce good results in an environment where vocal samples are well separated, background noise is kept to a minimum and an extensive set of training samples is available. However, that is not the case for this application. In addition to problems of separation and noise, there will be numerous unseen samples which we will wish to classify.

In general this section discusses the separation of an audio stream into samples which contain only one voice, and the classification of these samples. Separation of non–vocal sections of the audio stream into homogeneous samples is also desirable, however, the main focus here is on vocal similarity. The goal is to segment an audio stream into a set of *clean* samples, each of which contains only one voice, or contains a sample of non–vocal sound. Non–vocal samples will generally be discarded if they can be reliably separated from vocal samples.

There are two major aspects to the problem of separating an audio stream into clean or homogeneous samples. The first problem is to detect points at which the speaker changes. This is most commonly performed using silence as an indicator for segmentation points. In some applications, such as air traffic control dialogues [19], this performs well. However, in news video silence is not as effective at indicating changes in speaker or other sound changes. The second difficulty to aural segmentation is the detection of silence itself. Just as there are problems for indication of segmentation points by

silence, there are also characteristics of news broadcast that make detection of silence more difficult than it is in some other areas.

There are in fact a number of behaviours presented by anchor people that are used to maintain the narrative flow, which prevent accurate segmentation of sound samples. Anchor people often begin speaking before the audio from a piece of background footage has stopped, which aids narrative flow but makes it impossible to separate a voice in the footage from the anchor persons voice based on silence. Footage or advertisements following an anchor shot will usually fade in while the anchor person is still talking, once again leaving no clear silence between samples. In addition, the anchor person will generally start speaking before a cut from one shot to another, or will start speaking just after a cut, with sound from the previous sample continuing slightly past the cut. Due to this, most audio samples will contain more than one voice, regardless of the segmentation algorithm employed.

An additional aspect of this problem is that much of the audio track will contain noise as well as voice or other sounds of interest. Sound from news footage captured often contains noise of various forms, such as crowd noise in the background for field reports. This makes it difficult to recognise repetitions of a single voice, as variations in the associated noise can affect the attributes used for a similarity measure. This leads to a further consideration, which is that the length of audio sample must be long enough to allow stable statistical properties for the attributes chosen for a measure. Work by Gish, Sui and Rohlicek [6] has suggested that four seconds is a suitable minimum sample length for attributes of vocal samples to exhibit consistency, and this sample length is employed in this work.

There are three methods examined for segmentation of the sound track of a news video into four second samples. The basic method simply cuts the video every four seconds, starting from the initial frame (fixed interval method). The two other methods presented attempt intelligent segmentation based on detection of likely breaks in the narrative.

The simplest intelligent method uses cuts in the video to indicate likely breaks in sound. A cut is a point in the video where the vision changes from one camera shot to another. Cuts can be detected rapidly and reliably by a number of methods [13, 1], and would be expected to occur in the neighbourhood of changes in sound in many cases. As described earlier, this method frequently leads to a sample containing predominantly one voice, but also short segments of other voices which cross the cuts slightly. This is particularly true for interview or panel discussion formats, as the camera changes are often a reaction to a new speaker, rather than a scripted, and therefore known movement. An advantage of the cut method is that most commercials, which tend to have numerous very short shots, are excluded from classification.

The other intelligent method uses silence as an indicator for segmentation points. Detection of silence in news video is in itself problematic, as the level of energy which is defined as silence must be adaptive. Some earlier work assumes a sample of background noise will be present at the start of each clip, or that a consistent model of background noise will suffice [16, 22, 23]. This is unfortunately not always the case and background noise is often inconsistent within a single video stream. An example in news video is that background noise will differ between an anchor shot captured in a studio and a reporter captured on location in a crowded city scene. This is markedly different from environment such as telephone conversation [20] or air traffic control dialogues [19], with simple noise occurring when the speaker changes and during the speaker samples. Siu, Yu and Gish [19] note that due to such factors, in many applications a simple fixed length segmentation of audio will often give performance similar to silence detection.

3.1 Silence detection for aural segmentation

The algorithm used for silence detection in this paper attempts to separate vocal, from non–vocal sound in a low cost manner. The term silence in this section will refer to vocal silence, that is a section of the audio stream with no dominant vocal portion. The method employed is to detect sections of the audio stream which have dominant peaks in the spectrogram in the vocal region. Speech is largely restricted to the range of 100 Hz to 4500 Hz, and within this range the energy is often quite