An Affect-Based Video Retrieval System with Open Vocabulary Querying

Ching Hau-Chan[1] and Gareth J. F. Jones[1,2]

[1]Centre for Digital Video Processing
[2]Centre for Next Generation Localisation
School of Computing, Dublin City University, Dublin 9, Ireland
gjones@computing.dcu.ie

Abstract. Content-based video retrieval systems (CBVR) are creating new search and browse capabilities using metadata describing significant features of the data. An often overlooked aspect of human interpretation of multimedia data is the affective dimension. Incorporating affective information into multimedia metadata can potentially enable search using this alternative interpretation of multimedia content. Recent work has described methods to automatically assign affective labels to multimedia data using various approaches. However, the subjective and imprecise nature of affective labels makes it difficult to bridge the semantic gap between system-detected labels and user expression of information requirements in multimedia retrieval. We present a novel affect-based video retrieval system incorporating an open-vocabulary query stage based on WordNet enabling search using an unrestricted query vocabulary. The system performs automatic annotation of video data with labels of well defined affective terms. In retrieval annotated documents are ranked using the standard Okapi retrieval model based on open-vocabulary text queries. We present experimental results examining the behaviour of the system for retrieval of a collection of automatically annotated feature films of different genres. Our results indicate that affective annotation can potentially provide useful augmentation to more traditional objective content description in multimedia retrieval.

Keywords: affective computing, information retrieval, multimedia data, open vocabulary querying, automatic annotation

1 Introduction

The amount of professional and personal multimedia data in digital archives is currently increasing dramatically. With such large volumes of data becoming available, manually searching for a multimedia item from within a collection, which is already a time-consuming and tedious task, is becoming entirely impractical. The solution to this problem is to provide effective automated or semi-automated multimedia retrieval and browsing applications for users. This of course requires the data to be annotated with meaningful features to support user search. Unfortunately, it is unrealistic to expect all multimedia data to be

richly annotated manually therefore automated content analysis tools are vital to support subsequent retrieval and browsing.

According to [1] and [2] annotation can be differentiated into 3 levels as follows: labels at the lowest level (feature level) are primitive features such as shot cuts and camera motion, the next level is logical features (cognitive level) involving some degree of logical inference describing the content such as "red car below a tree" and finally the highest level (affective level) contains the most abstract features that involve some degree of subjectivity, such as "calm scene" or "funny face". Current multimedia retrieval systems are generally based on low-level feature-based similarity search. These systems are limited in terms of their interpretation of the content, but they are also difficult for non-expert users to work with since they typically want to retrieve information at the cognitive or affective level rather than working with low-level image features [3]. The difference between the low-level information extracted from multimedia data and the interpretation of the same data by the user in a given situation is identified as a *semantic gap* [4]. Developing methods to close the semantic gap to support more powerful and intuitive search of multimedia content is one of the ongoing research challenges in multimedia information retrieval. In complementary research, the field of *affective computing* focuses on the development of human-centered systems that can contribute to bridging this semantic gap [5]. For example, methods based on affective computing by allowing could enable users to query a system on a higher level of abstraction such as "find some exciting videos" instead of a low-level query describing features associated with the concept of "exciting" such as "rapid motion", "shot cuts" and "elevated audio energy".

In this paper we present work on a novel system designed to be a step towards providing this higher abstraction through an affect-based annotation of video content. The system automatically extracts a range of low-level audio and video features and then uses these to assign a set of affective verbal labels to the content. Video retrieval is then enabled using a system based on the Okapi retrieval model with an additional query pre-processing stage based on WordNet to provide open-vocabulary querying. Experimental retrieval results on a wide ranging collection of commercial movies show that affective annotation has the potential to augment existing multimedia search based on low-level objective descriptive features of the content.

This paper is organized as follows: Section 2 describes our affect extraction and labeling method for video data, Section 3 gives a summary description of the Okapi retrieval method used in our system, Section 4 presents our experimental investigations based on a movie collection, and Section 5 summarizes our conclusions and outlines possible directions for future work.

2 Affect Extraction and Labeling

One step towards bridging the semantic gap between user needs and detected low-level features is to combine these features to infer some form of higher-level features to which non-expert users can relate [6]. This method is favoured

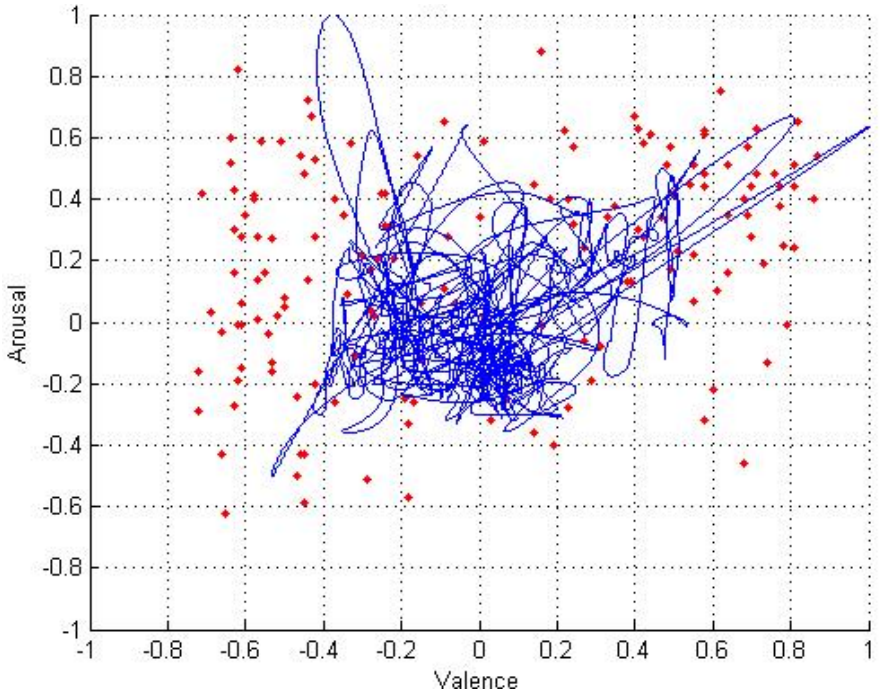

Fig. 1. Affect curve (line) mapped onto 151 emotional verbal labels (dots).

because it splits the semantic gap problem into two stages: mapping low-level features to immediate semantic concepts, and mapping these semantic concepts to user needs. As outlined in the previous section, existing research on multimedia content analysis and retrieval has concentrated largely on recognition of objective features describing what is observed in the content [7]. Such work includes scene segmentation [8][9], object detection [10], and sports highlight detection [11]. Our research aims to complement these objective features by using low-level features to describe an affective interpretation of the content. A detailed description of our annotation method is described in [12]. In this section we summarise our approach and in the following sections extend this existing work into an affect-based retrieval system.

The following subsections give a summary description of the annotation procedures for our affect-based retrieval system.

2.1 Modeling Valence and Arousal

The affective dimension of a video describes information of its emotive elements which are aimed at invoking certain emotional or affective states that humans can naturally identify. Therefore affective labels of the multimedia content relate to the affective states that the creator of the content is seeking to elicit in the

viewers. Including such labels in multimedia indexes would enable human users of multimedia retrieval systems to include expression of affective states as part of queries expressing their information needs. However, since affective states are subjective, even when presented with a set of affective labels individual users may select a different, but usually related, label to describe the same affective state that they associate with the desired multimedia content.

Research into human physical and cognitive aspects of emotion can help us to model affective features. One such model which is extremely useful in this context, is the Valence-Arousal-Dominance (VAD) emotion representation described by Russell and Mehrabian [13] which breaks emotions into 3 independent components. In the VAD model the 3 independent and bipolar dimensions represent all emotions experienced by humans. The three components are defined as follows:

- Valence: Measures the level of pleasure - displeasure being experienced. This ranges from a "positive" response associated with extreme happiness or ecstasy through to a "negative" response resulting from extreme pain or unhappiness.
- Arousal: This is a continuous measure of alertness ranging from one extreme of sleep through to intermediate states of drowsiness and alertness and finally frenzied excitement at the other end of the scale.
- Dominance: Dominance (or control) is a measure of a person's perceived control in a situation. It can range from feelings of a total lack of control or submissiveness to the other extreme of complete control or influence over their situation or environment. It has been observed that this dimension plays a limited role in affective analysis of video, and we follow previous work in concentrating only on valence and arousal [2].

In order to extract the affective content contained in video data, we first perform low-level feature extraction. Extended features are then combined to describe valence and arousal levels of the data as follows:

- Valence is modeled as a weighted sum of colour brightness and colour saturation from the visual stream and pitch from the audio stream with each weighting following findings reported in [14].
- Arousal is modeled as an equally weighted sum of global motion and shot cut rate from the visual stream, and energy from the audio stream as described in [2].

Each of the low-level features are subjected to a smoothing function and normalized in the range -1 and +1 to fulfill [2]'s comparability, compatibility, and smoothness criterion. The arousal and valence stream outputs of this process can be illustrated on a 2D VA plot showing an affect curve which plots valence and arousal against each other as illustrated in Figure 1. The affect curve illustrates the evolution of affective states contained in the data stream over time.

2.2 Verbal Labeling of the Affect Curve

To automatically annotate videos with affective labels, the affect curve is populated with emotional verbal labels from the findings of [13]. In this study averaged arousal, valence, and dominance values were assigned to 151 verbal labels by a group of human assessors. Figure 1 shows the 151 emotional verbal labels plotted as individual dots on the affect curve. The values are normalized in the range -1 to +1 so that the verbal labels can be mapped to the affect curve. It can be seen from Figure 1 that the labels are quite evenly distributed in the VA space enabling us to describe a wide range of emotions. Each point on the affect curve can be associated with the spatially closest label.

Automatically detecting the low-level visual and audio features of video data and processing it to obtain its affect curve allows us to annotate each frame of the video with a affective label. Taken over the duration of a multimedia document this generates a sequence of affect labels for the content. Therefore a simple frequency count of re-occurring labels can describe the major affective state(s) or emotional content of each part of the multimedia data. 151 verbal labels gives a quite fine level of labeling granularity, but since affective interpretation is subjective, we can also use the label stream to refer to alternative labels by choosing second or third closest labels. To study the granularity of labeling and the overall quality of affective labeling, we explored an additional two sets of verbal labels with coarser granularity. These consisted of 22 labels suggested by [15] and 6 labels based on word described in [16]. A similar approach was used in [17] using manual placement of 40 labels for the FEELTRACE system. A comparative investigation of these different labeling schemes is described in [12], as might be anticipated the overall conclusion was that there is a trade off between granularity of labels and reliability of individual labels. A larger number of labels mean a greater degree of expressivity, however inevitably the accuracy of label assignment will be reduced. Our experiments described later illustrate the need for a larger annotation vocabulary to support effective open-vocabulary search.

3 Information Retrieval and Document Matching

The classic information retrieval (IR) problem is to locate desired or relevant documents in response to a user information need expressed using a search query consisting of a number of words or search terms (which may be stemmed or otherwise pre-processed). Matching the search terms from the query with the terms within the documents then retrieves potentially relevant documents. Documents are ranked according to a query-document matching score measuring potential likelihood of document relevance. The user can then browse the retrieved documents in an attempt to satisfy their information need.

Using this principle with each visual and audio frame of the multimedia data with an affective label, the multimedia data can be thought of as a document containing re-occurring words where the frequency is directly related to the degree to which the affect associated with the label is present in the multimedia

item. This enables us to perform experiments into the retrieval of multimedia data from the perspective of affective content using a text IR model. Thus our system is based on entry of a text query which is used with an IR model to match the query with the system-detected affect labels to retrieve a ranked list of potentially relevant videos.

3.1 Okapi BM25 Information Retrieval Model

A number of IR algorithms have been developed which combine various factors to improve retrieval effectiveness. The effectiveness of these methods has been evaluated extensively using text test collections such as those introduced at the TREC evaluation workshops, see for example [18]. One of the most consistently effective methods for text retrieval is the Okapi BM25 IR model [19]. BM25 is a classical term weighting function. For a term i in a document j, the BM25 combined weight $CW(i,j)$ is:

$$CW(i,j) = \frac{CFW(i) \times TF(i,j) \times (K1+1)}{K1 \times ((1-b) + (b \times (NDL(j)))) + TF(i,j)}$$

where $K1$ and b are tuning constants. $CFW(i) = \log(N/n(i))$ is the collection frequency weight where N is the total number of documents in the collection and $n(i)$ is the total number of documents containing term i. $TF(i,j)$ is the frequency of term i in document j. The BM25 formula overall ensures that the effect of term frequency is not too strong, and for a term occurring once in a document of average length that the weight reduces to a function of $CFW(i)$ for a document of average length. The overall matching score for a document j is simply the sum of the weights of the query terms present in the document. Documents are ranked in descending order of their matching score, for presentation to the user. The tuning constant $K1$ modifies the extent of the influence of term frequency. The constant b, which ranges between 0 and 1, modifies the effect of document length. If $b = 1$ the assumption is that the documents are long simply because they are repetitive, while if $b = 0$ the assumption is that they are long because they are multi-topic. Thus setting b towards 1, such as $b = 0.75$ will reduce the effect of term frequency on the grounds that it is primarily attributable to verbosity. If $b = 0$, there is no length adjustment effect, so greater length counts for more, on the assumption that it is not predominantly attributable to verbosity.

Since Okapi BM25 has been shown to be effective in many retrieval settings, we adopt it in our affect-based retrieval system.

3.2 Open-Vocabulary Query

In standard text IR, documents and queries both use an open vocabulary. For a well constructed IR system, the success of an IR system relies on there being a good match between words appearing in relevant documents and the submitted search request.

The affect labeling described in section 2.2 is limited to 151 labels. While the 151 labels cover a wide range of possible affective states, it is likely that users will often use query words that are not part of this list. This mismatch between user query and detected affect labels in the system will mean that the relevant documents may often not be retrieved, or be retrieved unreliably at a reduced rank. In order to address this problem we use a novel solution to enable open-vocabulary querying. In this method a measure of relatedness is calculated between each query word entered by the user and the list of affective labels used by the system. This ensures that even if the query word is not one of the annotated multimedia labels, the system is able to produce a ranked list of closest match multimedia documents for the user.

We use WordNet [20], a freely available lexical database/dictionary consisting of nouns, verbs, adjectives, and adverbs organized into a network of related concepts called synonym sets to provide our open-vocabulary word relatedness scoring. This provides us with a tool to measure semantic relatedness or similarity between different words. Our system uses WordNet::Similarity, a freely available Perl module that implements the similarity and relatedness measures in WordNet [20].

A measure of similarity quantifies how much two concepts (or words) are alike based on the information contained in WordNet's relations or hierarchy. Due to the organization of words into synonym sets, two words can be said to be similar if counting the distance of the relation or hierarchy results in a small distance. Extending this further, a similarity measure can be derived by counting the path lengths from one word to another, utilizing the *is-a* relation. [21] presents an algorithm to find the shortest path between two concepts and scales this value by the maximum path length D in the *is-a* hierarchy in which they occur. A different take on a similarity measure is proposed by [22], which uses knowledge of a corpus to derive a similarity measure. Their similarity measure is guided by the intuition that similarity between a pair of words may be judged by the extent to which they share information.

However *is-a* relations in WordNet do not cross part-of-speech boundaries, so such measures are limited to judging relationships between noun pairs and verb pairs. The system presented in this paper relies on adjective words that describe emotions such as "happy", "joyful" and "sad" in order to describe affective states. Adjectives and adverbs in WordNet are not organized into *is-a* hierarchies, but can still be related through antonyms and *part-of* relations, called measures of relatedness. For example, a "wheel" has a *part-of* relationship with a "car", and "happy" is the opposite of "unhappy".

We use the measure of relatedness proposed by [23], where the idea of semantic relatedness is that two words are semantically close if their WordNet synonym sets are connected by a path that is not too long and does not change direction too often. For every query label, we use WordNet::Similarity to compare it to the 151 affective labels. If a label is highly related it will score higher, while query words that are found in the list of 151 affect labels are given the maximum score.

3.3 Query Processing Strategies

We explored six different query processing strategies in a known-item search task discussed in the next section. The first four strategies use the open-vocabulary query processing to map the user query words to the annotation system's 151 affective labels to form a final query that is fed into the IR system for retrieval. The four strategies are referred to as: "Unweighted full expansion", "Weighted full expansion", "Unweighted best expansion", and "Weighted best expansion". The remaining two strategies are referred to as: "Reweighted" and "Bypass".

- Full expansion means that all labels that have relatedness scores above 0 for each user-supplied query word were fed into the IR model. This in effect fully expands the user query to the all available labels, hence "full expansion".
- Best expansion means that only affect labels with the highest (best) relatedness score for each user query word were input to the IR model. Sometimes a query label will have 2 labels with identical highest relatedness scores, in these cases both labels were used.
- Unweighted or Weighted determines whether the final expanded query input to the IR model was weighted according to the relatedness scores calculated from WordNet. For the Unweighted strategy, every word included in the query was said to be equally important. For the Weighted strategy, each query word's relatedness score was multiplied by the combined weight $CW(i, j)$ of the IR model to generate the ranked retrieval list.

The following equations show the mathematical formulae for the modified BM25 weights for the full expansion strategies, if the relatedness score $rel_{HS}(i_R, i) \neq 0$, where i_R is the relevant term.

- Unweighted full expansion:

$$CW_{UW}(i, j) = \sum_{i_R} CW(i_R, j)$$

- Weighted full expansion:

$$CW_W(i, j) = \sum_{i_R} rel_{HS}(i_R, j) \times CW(i_R, j)$$

The following strategies are applied for the best expansion strategies if the relatedness score $rel_{HS}(i_R, i) = \max$, where i_R is the scored as the most related term.

- Unweighted best expansion:

$$CW_{BUW}(i, j) = \sum_{\substack{i_R \\ rel_{HS}(i_R, i) = \max}} CW(i_R, j)$$

– Weighted best expansion:

$$CW_{BW}(i,j) = \sum_{\substack{i_R \\ rel_{HS}(i_R,i)=\max}} rel_{HS}(i_R,j) \times CW(i_R,j)$$

The fifth strategy is a modification of the "Weighted best expansion" strategy where the relatedness score was first raised by an exponent value X before it was multiplied with the combined weights of the IR model. This gives the higher range of the relatedness scores heavier weights and emphasis, if the relatedness score $rel_{HS}(i_R,j) = \max$, where i_R is the relevant term and X is the exponent value.

– Reweighted best expansion:

$$CW_{BRW}(i,j) = \sum_{i_R} rel_{HS}(i_R,j)^X \times CW(i_R,j)$$

The sixth strategy called "Bypass" is to bypass the open vocabulary query mapping of the system and directly feed the queries into the IR model to generate a ranked list, therefore any query words not found in the system's list of affect labels are simply ignored. The following combined weight is obtained if query term i is found in the list of affective labels, where i_R is the relevant term.

– Bypass:

$$CW_B(i,j) = \sum_{i_R} CW(i_R,j)$$

In each case the standard $CW(i_R,j)$ weight in the BM25 function is replaced by the relevant modified version.

4 Experimental Investigation

The experimental investigation presented here explores the behaviour and potential of our affect-based retrieval system for searching a collection consisting of a variety of commercial Hollywood movies. Movies of this type represent a compelling source of video data for an affect-based system due to the richness of their emotional content. They are additionally suitable for our initial study of affect-based search since emotional content is much more pronounced in movies than in other video material, making it easier to determine what emotions a movie is trying to project. A total of 39 movies were processed covering a wide range of genres from action movies to comedy and horror movies. This amounted to approximately 80 hours of data comparable to video evaluation campaigns such as TRECVid [24]. Each movie was broken up into 5 minute clips, giving a total of 939 film clips.

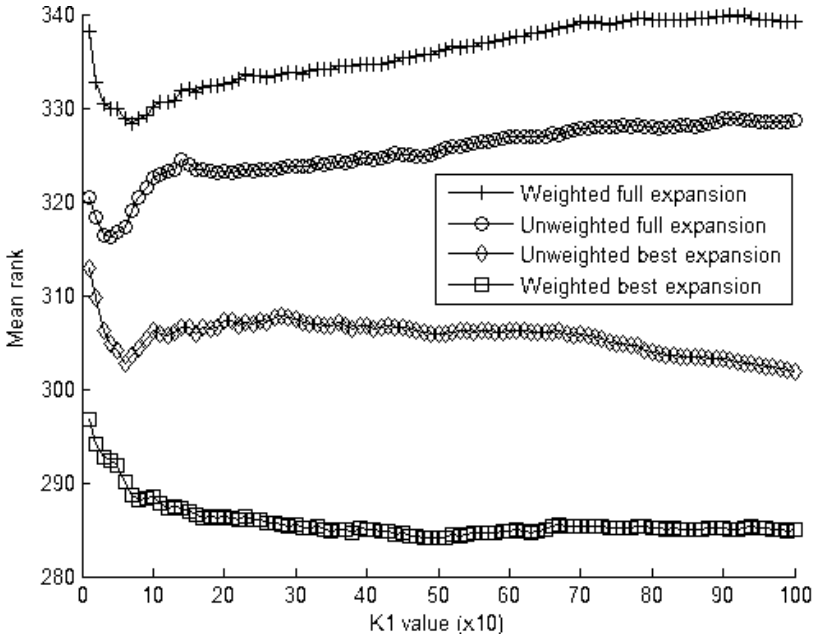

Fig. 2. Mean rankings for the 4 query processing strategies.

4.1 Known-Item Search Task

A known-item search task was performed to measure the system's effectiveness at retrieving and locating the original film clip described by a text query, from the database of films using a textual description of that particular clip's emotional content labelled using our affect label assignment system.

In order to generate the test set, 8 volunteers were each randomly assigned a unique set of 5 film clips. They viewed each of these clips and then created an open-vocabulary affective textual description of it. These descriptions were collected together as a set of 40 search queries for the known-item search task.Depending on the query processing strategy used, the user query was mapped to the system's affective list of 151 labels, and these labels used as the final query. The system generated a ranked list of clips from the database of movies using the Okapi BM25 IR model. The ranked list sorted the clips in order of relevance. From this ranked list, the original clip's position on this list was identified. The higher its position in the list, the better the retrieval performance.

4.2 Experimental Results

Figure 2 shows how different $K1$ values affect the mean rankings of the relevant film clip for the first 4 query processing strategies. A thousand runs were performed using the system to automatically calculate the mean rank for the 40 queries. The b value was set to 0 in all cases. There is a noticeable dip in mean

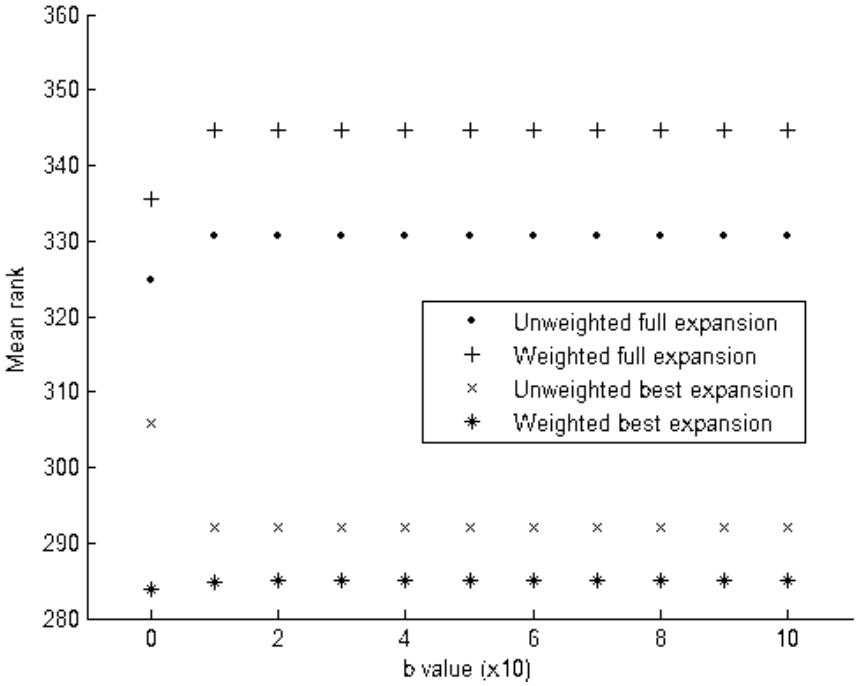

Fig. 3. Mean rankings for the 4 query processing strategies.

Table 1. Number of returned results for the query processing strategies.

Total queries: 40	Unweighted full exp.	Weighted full exp.	Unweighted best exp.	Weighted best exp.	Bypass
Number of returned results	32	32	26	26	3
Recall	0.8	0.8	0.65	0.65	0.08

rank for all strategies when the $K1$ value reaches 51. It can be observed that the best mean rank achieved by the "Weighted best expansion" strategy is when $K1$ is at the value 474.

Figure 3 shows that the best mean ranks of relevant items was achieved when the b value was set to 0, except for the "Unweighted best expansion" strategy where the mean rank of the relevant clips were degraded when b was set to 0. When the b values changes from 0 to 1, the mean ranks does not appear to change. Closer inspection of the values reveal that the mean ranks do change with different b values, but that the variation was too small to be noticeable on the graph. The b value in the BM25 model relates to the topical structure of documents and document length, since all documents were of the same length with similar topical structuring, it is unsurprising that b is not a significant component in retrieval effectiveness for this task.

Table 2. Comparison of the ranks for the query processing strategies

Queries relevant result for "Bypass"	Unweighted full exp.	Weighted full exp.	Unweighted best exp.	Weighted best exp.	Bypass
Query 4	346	451	346	451	58
Query 14	301	211	182	94	6
Query 36	51	13	12	14	7

Table 3. Mean rank and mean-reciprocal-rank (MRR) for the query different expansion strategies for 26 queries retrieving relevant items with Best Expansion methods.

	Unweighted full exp.	Weighted full exp.	Unweighted best exp.	Weighted best exp.	Reweighted best exp. (16)
Mean rank	301.9	284.8	307.8	308.5	271.0
MRR	0.027	0.019	0.022	0.0221	0.025

Table 1 shows the number of queries for which the system successfully retrieved the target clip using the 5 query processing strategies. The two "full expansion" strategies retrieved 32 clips out of 40, giving the best recall rate of 0.8. The "Bypass" strategy only retrieved 3 clips, with the lowest recall rate of 0.08. This indicates that as the user word query is mapped to an increasingly smaller number of labels to form the final query, the recall rate drops. Note that the failure to retrieve a clip at any rank indicates that none of the query words were contained in the label from the affective label list automatically assigned to the target clip in the analysis stage.

Table 2 shows the individual retrieved rank results for the three queries for which the relevant item was retrieved by the "Bypass" strategy and their ranks across the different strategies. It can be observed that the "Bypass" strategy achieved the best ranks for all three of these queries. These results illustrate that exact matches with the affect label list in the query can be very effective for retrieval. The difficulty with limiting the query vocabulary to only these labels is that users are likely to find this list of words constraining and difficult to use in describing their information needs. Hence expansion to alternative labels is needed for good recall levels, but at the cost of greatly degraded average rank at which relevant items are retrieved.

Mean-reciprocal-rank (MRR) is calculated as the mean of the reciprocal of the rank of retrieved relevant items, using 0 for queries for which the known-item is not retrieved. Compared to the mean rank, the MRR has the effect of not punishing a system excessively for retrieval of individual items at very low rank. Thus, it gives a better indication of the average performance across a query set. The nearer the MRR is to 1.0, the better the system is performing on average. Table 3 shows the mean rank and MRR results for the 4 expansion strategies for the 26 topics for which the "best expansion" strategy retrieves the relevant item, these are a subset of the 32 queries for which the relevant item is

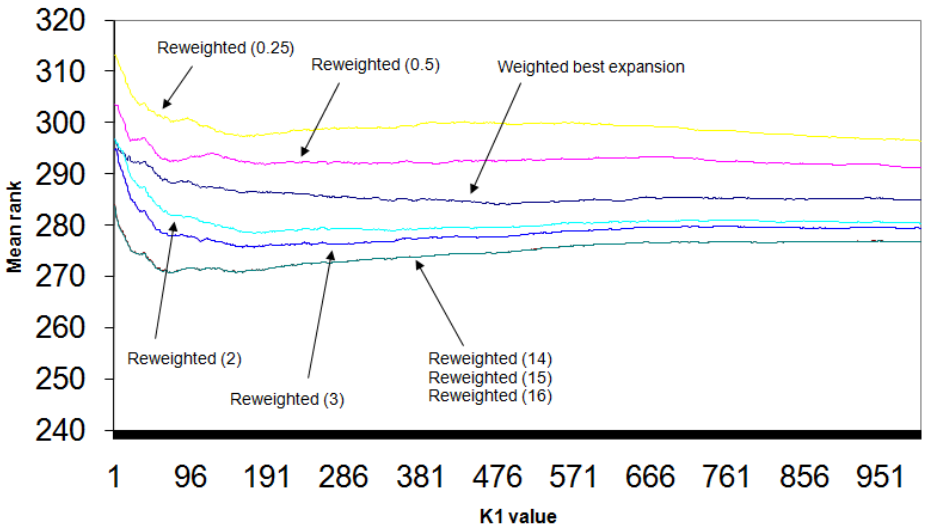

Fig. 4. Mean rankings for different exponent values for the query processing strategies.

retrieved by "full expansion". This subset was used to enable a direct comparison of results for the different strategies. It can be seen that the best MRR is given by "Unweighted full expansion". Examination of individual results shows that with "Unweighted full expansion" for a small number of queries the relevant item is retrieved at a much higher rank than with the other strategies, although on average it performs less well than the other strategies, leading to its better MRR and worse mean rank effectiveness than the "Best" expansion strategies.

Figure 4 shows the mean ranks for the re-weighted (fifth) "best expansion" strategy, where the numerical value enclosed in round brackets is the exponent value. It can be seen that as the value of the X parameter increases, the mean ranks are improved over the original strategy. The mean ranks improvement stops when the exponent value reaches $X = 14$, as can be seen in the figure, the lines ($X = 15$ and $X = 16$) are almost identical to the line for $X = 14$. In addition, the best $K1$ value for the re-weighted strategies was found to be 71. The final column of Table 3 shows the mean rank values and MRR for the re-weighted "best expansion" for $k = 71$. It can be seen that the exponential function gives improvement in both the mean ranks and MRR values. This is the best mean rank result, while the MRR result is still slightly less than that achieved with "unweighted full expansion" for the reasons given earlier.

5 Conclusions and Further Work

The results of experiments reported in this paper show that video clips which had been described by users using text queries can be retrieved with a measure of consistency for affect-based user queries. While the WordNet expansion is

shown to improve recall, it is clear that where the affect label vocabulary covers the query words that retrieval effectiveness is better. Thus, it would appear that rather than seeking an improvement through increased technical sophistication, the most effective strategy would be to generally increase the affective label set that can be placed on the VA plot. While in the past such an endeavour would have been costly and difficult to arrange, crowdsourcing methods such as Mechanical Turk[1] could potentially make it relatively straightforward to gather valence and arousal values for very many words averaged across a large number of people. A more sophisticated method of assigning labels to video data would then be required since the single assignment of a label based on VA proximity will not be sufficiently accurate, labels might be clustered to an average location or perhaps multiple labels might be assigned based on some proximity measure.

Also since affect is a subjective interpretation of an important, but limited, dimension in describing multimedia content, we believe that affect-based annotation is more likely to be used most effectively to augment existing objective multimedia content retrieval systems, rather than to be used independently. Further work is planned to explore how this alternative dimension of indexing and search can be incorporated into existing multimedia retrieval systems.

6 Acknowledgements

This research is partially supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) at Dublin City University.

References

1. Eakins, J.P.: Automatic image content retrieval - are we getting anywhere?, Proceedings of the 3rd International Conference on Electronic Library and Visual Information Research, De Montfort University, Milton Keynes, U.K., pages 123-135 (1996)
2. Hanjalic A., and Qun Xu, L.: Affective Video Content Representation and Modeling, IEEE Transactions on Multimedia, 7(1):143-154, 2005.
3. Lee, H., Smeaton, A.F., McCann, P., Murphy, N., O'Connor, N., Marlow, S.: Fschlr on a PDA: A Handheld User Interface to a Video Indexing, Browsing, and Playback System, ERCIM Workshop User Interfaces for All, Florence, Italy (2000)
4. Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1349-1380 (2000)
5. Lew, M. S., Sebe, N., Djeraba, C., Jain, R.: Content-based Multimedia Information Retrieval: State of the Art and Challenges, ACM Transactions on Multimedia Computing, Communications and Applications, 2(1):1-19 (2006)
6. Hauptmann, A.G.: Lessons for the Future from a Decade of Informedia Video Analysis Research, Proceedings of the 4th International Conference on Image and Video Retrieval (CIVR 2005), Singapore, pages 1-10 (2005)

[1] `https://www.mturk.com`

7. Hauptmann, A.G., Christel, M.G.: Successful Approaches in the TREC Video Retrieval Evaluations, , Proceedings of the Twelfth ACM International Conference on Multimedia 2004, New York, NY, USA, pages 668-675 (2004)
8. Zhang, T., Jay Kuo, C.-C.: Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing, Kluwer Academic Publishers (2001)
9. Zhai, Y., Shah M., Rasheed, Z.: A Framework for Semantic Classification of Scenes using Finite State Machines, Proceedings of the Conference for Image and Video Retrieval, Dublin, Ireland, pages 279-288 (2004)
10. Browne, P., Smeaton, A. F.: Video Information Retrieval Using Objects and Ostensive Relevance Feedback, ACM Symposium on Applied Computing, Nicosia, Cyprus, pages 1084-1090 (2004)
11. Sadlier, D., and O'Connor, N.: Event Detection based on Generic Characteristics of Field Sports, IEEE International Conference on Multimedia and Expo (ICME 2005), Amsterdam, The Netherlands, pages 759-762 (2005)
12. Jones, G. J. F., Chan, C. H.: Affect-Based Indexing for Multimedia Data, Multimedia Information Extraction, ed. M.T. Maybury, IEEE Computer Society Press (2011)
13. Russell, J. A., Mehrabian, A.: Evidence for a Three-Factor Theory of Emotions, Journal of Research in Personality, 11:273-294 (1977)
14. de Kok, I.: A Model for Valence Using a Color Component in Affective Video Content Analysis, Proceedings of the Fourth Twente Student Conference on IT, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands (2006)
15. Salway, A., Graham, M.: Extracting Information about Emotions in Films, Proceedings of the Eleventh ACM International Conference on Multimedia 2003, Berkeley, CA, USA, pages 299-302 (2003)
16. Ekman, P., Friesen, W.V.: Facial Action Coding System, Consulting Psychologists Press Inc, Palo Alto, California, USA (1978)
17. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schrder, M.: 'FEELTRACE' : An Instrument for Recording Perceived Emotion in Real Time, Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, Belfast, U.K., pages 19-24 (2000)
18. Harman, D. K.: The Fifth Text Retrieval Conference (TREC-5), National Institute of Standards and Technology, Gaithersburg, MD (1997)
19. Robertson, S.E., Spärck Jones, K.: Simple, proven approaches to text retrieval, Technical Report, TR356, Cambridge University Computer Laboratory (1997)
20. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet::Similarity - Measuring the Relatedness of Concepts, Proceedings of the 19th National Conference on Artificial Intelligence (AAAI 2004), San Jose, CA, USA (2004)
21. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification, WordNet: An electronic lexical database, MIT Press, 265-283 (1998)
22. Resnik, P.: Using information content to evaluate semantic similarity, Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, 488-453 (1995)
23. Hirst, G., St-Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms, Wordnet: An Electronic Lexical Database, MIT Press, 305-332 (1998)
24. Smeaton, A. F., Over, P.: Kraaij, W., Evaluation campaigns and TRECVid, MIR '06: Proceedings of the Eighth ACM International Workshop on Multimedia Information Retrieval, Santa Barbara, CA, USA , pages 321-330 (2006)