A Human-based Technique for Measuring Video Data Similarity

Waleed E. Farag*  Hussein Abdel-Wahab**

*Computer Science Department, Indiana Univ. of Pennsylvania, Indiana, PA 15705, USA
Email: farag@iup.edu
**Computer Science Department, Old Dominion Univ., Norfolk, VA 23529, USA
Email: wahab@cs.odu.edu

Abstract

The increasing use of multimedia streams nowadays necessitates the development of efficient and effective methodologies for manipulating databases storing them. Moreover, content-based access to multimedia databases requires in its retrieval stage to effectively assess the similarity of video data. This work proposes a new technique for measuring video data similarity that attempts to model some of the factors that reflect human notion in evaluating video data similarity. This model presents one step towards designing intelligent content-based video retrieval systems capable of measuring the similarity among video clips in a way similar to what humans do. The performance of the proposed model was tested in terms of recall and precision of the retrieved results where the system yielded very satisfactory values of recall and precision under various testing scenarios.

1. Introduction

The basic objective of any automated video indexing system [2] is to provide the user with easy-to-use and effective mechanisms to access the required information. For that reason, the success of a content-based video access system is mainly measured by the effectiveness of its retrieval phase. The general query model adapted by almost all multimedia retrieval systems is the QBE (Query By Example) [24]. In this model, the user submits a query in the form of an image or a video clip and asks the system to retrieve similar data.

Upon the reception of the submitted query, the retrieval stage analyzes it to extract a set of features then performs the task of similarity matching. In the latter task, the query-extracted features are compared to the features stored into the metadata then matches are sorted and displayed back to the user based on how close a hit is to the input query. A central issue here is how the similarity matching operations are performed and based on what criteria. This central theme has a crucial impact on the effectiveness and applicability of the retrieval system.

In this paper, we discuss the design of the retrieval stage of our VCR (Video Content-based Retrieval) system and shed the light on one of its distinguishing characteristics. Our system attempts to make its comparison decisions based on modeling the way humans perform similarity matching of video data. This is achieved by using a number of factors reflecting how humans perceive media similarity [13]; thus, the proposed model overcomes the shortcomings of other approaches. The VCR system has, in addition to the retrieval stage discussed in this paper, three other modules namely, shot boundary detection [6], [7], key frames selection [8], and the indexing module [9].

The database population phase is performed as an offline activity and it outputs a set of metadata with each element representing one of the clips in the video archive. The retrieval system provides the user with an easy-to-use and effective interface through which the user can query and browse through the video library. To achieve this task, the retrieval system performs the following steps:

- Shot boundary detection and key frames selection are performed if the query is a video clip then features extraction is done next, otherwise features extraction is performed directly on the JPEG [15] query image.
- Measuring the similarity among the extracted features and those stored into the metadata.
- Results are returned back to the user.

The remainder of this paper is organized as follows. We review briefly in section 2 a number of related approaches for designing video retrieval systems. The details of the retrieval system are introduced in section 3 along with the filtering stage used to speed up the search process. In section 4, the proposed similarity matching model is expounded including video shot similarity definition and the adapted similarity measuring factors. Performance evaluation of the proposed retrieval system is given in section 5 followed by conclusion in section 6.
2. Literature Review

The worth of any information repository is only determined by how effective one can retrieve the required information [1]. For that reason, many researchers started to be aware of the significance of providing effective tools for accessing video databases and a review to some of these techniques is given below.

In the context of the browsing capability, some researchers proposed the integration between the human and the computer to improve the performance of the retrieval stage. In [14] a system is proposed that allows the user to define video objects on multiple frames then the system can interpolate the video object contours in every frame. Another video browsing system is presented in [20], [21] where comic book style summaries are used to provide fast overviews of the video content. One other prototype retrieval system that supports 3D images, videos, and music retrieval is presented in [12].

One technique proposed in [3] uses the metadata derived from clip links and the visual content of the clip to measure video similarity. Color histograms are used to represent the visual content in conjunction with a pruning measure video similarity. Color histograms are used to represent the visual content in conjunction with a pruning measure video similarity. An extension to the video signature technique just described that clusters stored data is introduced in [4]. A different retrieval approach uses time-alignment constraints to measure the similarity and dissimilarity of temporal documents [23]. In [19], a framework is proposed to measure video similarity. It employs different comparison resolutions for different phases of video search and uses color histograms.

A powerful concept to improve searching multimedia databases is called relevance feedback [17], [22], [25]. In this technique, the user can associate a score to each of the returned hits and these scores are used to direct the following search phase and improve its results. One example introduced in [25] defines the relevance feedback as a biased classification problem and uses linear/non-linear bias discriminant analysis to solve it.

From this quick survey of the current approaches, we can observe that an important issue has been overlooked by almost all the above techniques. This issue can be stated as “similarity matching has significance only if it can emulate what humans do” [18]. Our belief in the utmost importance of the above phrase motivates us to propose a novel human-based model to measure the similarity of video data.

3. The Retrieval System

The first component of the proposed system is the user interface whose main functions are:

- To provide users with an effective visualization tool to browse through the results and to select any one to be played or to be used as a new search example.
- To provide users with an effective visualization tool to browse through the results and to select any one to be played or to be used as a new search example.

If the QBE assumption is not valid (there is no example query), the user can start by browsing the database then selecting an item to be used as a query example.

After deriving the query extracted features, the system can not compare them with all the clips in the database because the time to retrieve the required video clip will be directly proportional to the number of stored clips. To tackle this issue, a two-stage search procedure is proposed to improve the efficiency and scalability of the search process, these stages are:

- The filtering stage: Where the accuracy of the matching operation is not high.
- The actual comparison stage: Where detailed matching operations are performed.

3.1. The Filtering Algorithm

The algorithm we proposed to implement the filtering stage is a simple but effective way to reduce the size of the search space. It works only if the input query is a video clip and it does not apply to image queries. The basic concept is to derive a signature for each shot and compare these signatures (with some tolerance) at the beginning of the search process. If the signature of a query shot is approximately similar to a metadata shot signature, the database video clip containing this shot is considered relevant to that query shot. That database clip is now a candidate for the second search stage. One assumption we postulate here is that if one shot of a database clip is relevant to one of the query shots, the whole database clip will be elected as one input to the accurate comparison stage, the second one. The signature we used to represent each shot is the relative distance between selected key frames. The experimental results section will assess the algorithm performance.

3.2. Browsing and User Interface

We designed an easy-to-use user interface with two major goals in mind. The first one is to allow the user to navigate the video archive to select a query. Our second goal is to provide the user with an effective visualization tool to browse through the returned search results. A snapshot of part of the user interface displaying some of the retrieved clips is depicted in Figure 1.

The proposed browsing environment does worth some explanation at this point. We propose a visual browsing approach where each shot is represented (in the display interface) by its first key frame. The initial display of a set of shots will be their representative key frames (one for each shot). The results of a query will be displayed using the key frames organization just described and the same
structure will allow the user to navigate the returned hits. This interface enables the user to select any of the returned clips to be used as a new query and this process can be repeated many times.

4. The Similarity Matching Model

The outputs of the filtering stage will be used as inputs to the second search stage that performs accurate similarity matching based on the contents of the query and those stored into the database. Before discussing the details of the proposed similarity model, let us define the three types of input queries the system can accept:

- A query image: In this case, neither the filtering stage nor the human-based similarity factors are applicable because the image has no temporal dimension.
- A one-shot video clip: This is a special case of the next type and in these two types both the filtering stage and the similarity matching model are applicable.
- Multi-shot video clip: This is the general case where the input is a video clip having more that one shot. The retrieval system will analyze the clip, extract its features, and perform the similarity matching operations.

In order to lay the foundation of our similarity matching model, a number of assumptions are listed first:

- The similarity of video data (clip-to-clip) is based on the similarity of their constituent shots.
- Two shots are not relevant, if the query signature is longer than other signature.
- A database clips is relevant, if one query shot is relevant to any of its shots.
- The query clip is usually much smaller than the average length of database clips.

The results of submitting a video clip as a search example is divided into two levels. The first one is the query overall similarity level which lists similar database clips. In the second level, the system displays a list of similar shots to each shot of the input query and this gives the user much more detailed results based on the similarity of individual shots to help fickle users [16] in their decisions.

4.1. Shot Similarity Definition

A shot is a sequence of frames so we need to formulate first frames similarity. In our model, the similarity between two video frames is defined based on their visual content where color and texture are used as visual content representative features. Color similarity is measured using the normalized histogram intersection while texture similarity is calculated using a Gabor wavelet transform. We use equation (1) to measure the overall similarity between two frames \( f_1 \) and \( f_2 \) where \( S_c \) (color similarity) is defined in equation (2). A query frame histogram \( H_q(i) \) is scaled before applying equation (2) to filter out variations in video clips dimensions. We define \( S_t \) (texture similarity) in details in [9].

\[
Sim(f_1, f_2) = 0.5 * S_c + 0.5 * S_t \quad (1)
\]

\[
S_c = \left[ \frac{\sum_{i=1}^{64} \text{Min}(H_1(i), H_2(i))}{\sum_{i=1}^{64} H_1(i)} \right] / n1 \quad (2)
\]

Suppose we have two shots \( S1 \) and \( S2 \) each has \( n1 \) and \( n2 \) frames respectively. We measure the similarity between these shots by measuring the similarity between every frame in \( S1 \) with every frame in \( S2 \) and form what we called the similarity matrix that has a dimension of \( n1x n2 \). For the ith row of the similarity matrix the largest element value represents the closest frame in shot \( S2 \) that is most similar to the ith frame in shot \( S1 \) and vice versa. After forming that matrix, equation (3) is used to measure shot similarity.

\[
Sim(S1, S2) = \left[ \sum_{i=1}^{n1} \frac{\text{MR}_{ij}(S_{i,j})}{\text{MC}_{ij}(S_{i,j})} + \sum_{j=1}^{n2} \frac{\text{MC}_{ij}(S_{i,j})}{\text{MR}_{ij}(S_{i,j})} \right] / (n1 + n2) \quad (3)
\]

Where: \( \text{MR}_{ij}(S_{i,j}) \) / \( \text{MC}_{ij}(S_{i,j}) \) is the element with the maximum value in the i/j row/col respectively.

\( n1/n2 \): the number of rows/columns in the similarity matrix.

Equation (3) is applied upon the selected key frames to improve efficiency and avoid redundant operations.

4.2. Human-based Similarity Factors

Our similarity model attempts to emulate the way humans perceive the similarity of video material. This was achieved by integrating into the similarity measuring formula a number of factors that most probably humans use to perceive video similarity. These factors are:

- **The visual similarity**: Usually humans determine the similarity of video data based on their visual characteristics such as color, texture, shape, etc. For instance, two images with the same colors are usually judged as being similar.
- **The rate of playing the video**: Humans tend also to be affected by the rate at which frames are displayed and they use this factor in determining video similarity.
The time period of the shot: The more the periods of video shots coincide, the more they are similar to human perception.

The order of the shots in a video clip: Humans often give higher similarity scores to video clips that have the same ordering of corresponding shots.

We included all these factors into one similarity matching formula given in equation (4).

\[
Sim(S1, S2) = W_1 \cdot S_1 + W_2 \cdot D_R + W_3 \cdot F_R
\]

(4)

Where: \( S_i \): The visual similarity, \( D_R \): Shot duration ratio, \( F_R \): Video frame rate ratio, \( S_i(d) \): Time duration of the ith shot, \( S_i(r) \): Frame rate of the ith shot, and \( W_1, W_2, \) and \( W_3 \): Relative weights.

There are three parameter weights in equation (4), namely, \( W_1, W_2, \) and \( W_3 \) that give indication on how important a factor is over the others. For example, stressing the importance of the visual similarity factor is achieved by increasing the value of its associated weight \( (W_1) \). The question of how to select the best values of these weights has been addressed first by exploring randomly selected combinations or using an optimization algorithm such as the Genetic Algorithm [5], [10]. These trials resulted in the following conclusion. Selecting fixed similarity criteria or evaluation function(s) by one or even few individuals do not necessary reflect the opinions of other users of the system. Consequently, we decided to give the user the ability to express his/her real need by allowing these parameters to be adjusted by the user. Three easy-to-use sliders are supplied into the interface (see Figure 1) to achieve this goal.

To reflect the effect of the order factor, the overall similarity level checks if the shots in the database clip have the same temporal order as those shots in the query clip. Although this may restrict the candidates to the overall similarity set to clips that have the same temporal order of shots as the query clip, we still have a finer level of similarity that is based on individual query shots which captures other aspects of similarity as discussed before.

5. Performance Evaluation

To evaluate the performance of our system we start by using the standard Precision \((P)\) and Recall \((R)\) metrics [11] defined in equations (7) and (8)

\[
R = A / (A + C)
\]

(7)

\[
P = A / (A + B)
\]

(8)

\( A \): correctly retrieved, \( B \): incorrectly retrieved, \( C \): missed

For efficiency reason, all of the following experiments use only the color feature in calculating the visual similarity of video data. Before discussing the details of the experimental results, it is important to note that the values of both recall and precision depend on the number of allowed shots to be retrieved as a response to the query. Thus, during the following experiments this number will be changed to measure the values of recall and precision. One parameter that needs to be determined before calculating recall and precision is the set of relevant shots to a particular query known as the ground truth. This ground truth set is determined manually, by a human observer, before submitting a query to the system. Because of being a manual process, the selection of the ground truth is performed on a subset of our database.

5.1. Experimental Results Using Database Queries

We started by choosing the ground truth for five shots with diverse contents that are parts of the database but they are not shown here for space constraints. Figure 2 shows part of the retrieval interface displaying the first ten hits resulted from submitting shot 9 of smg-npa-3 clip. All the returned shots are shots containing the same character pictured in the query and the first hit is the query shot itself. Moreover, some of the returned shots belong to the same query clip while the others are shots from a different clip. The performance of the retrieval system is very good in this query in which all relevant shots have been retrieved without any misses or false alarms.

To measure the recall and precision of the system, five shots were submitted as queries while changing the returned shots number from 5 to 20. The results of this procedure for the action-movie clip are shown in Table 1.

![Figure 2. First ten hits for shot 9 of the smg-npa-3 clip.](image)

Table 1. Recall and Precision as Functions of Returned Shots Number for Shot 2 of action-movie

<table>
<thead>
<tr>
<th>No. of returned shots</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>R</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>12</td>
<td>0.294</td>
<td>1.0</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>1</td>
<td>8</td>
<td>0.529</td>
<td>0.9</td>
</tr>
<tr>
<td>15</td>
<td>13</td>
<td>2</td>
<td>4</td>
<td>0.761</td>
<td>0.867</td>
</tr>
<tr>
<td>20</td>
<td>17</td>
<td>3</td>
<td>0</td>
<td>1.0</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Proceedings of the Eighth IEEE International Symposium on Computers and Communication (ISCC’03) 1530-1346/03 $17.00 © 2003 IEEE
A plot for the obtained recall results is depicted in Figure 3 where the five shots are named a, b, c, d, and e respectively. One can observe the predicted increase in the recall value with the increase in the number of returned shots and as soon as the latter number exceeds the number of relevant shots the recall value reaches 1.0 for all the shots considered in these figures. Another remark is that the recall curves for shots b, d, and e coincide, showing that the system exhibits relatively steady performance over different query contents. Similarly, the accuracy of the system is very good and in some cases its value reaches one regardless the number of the returned shot. Both recall and precision depend on the number of returned shots. To increase recall, more shots have to be retrieved, which will in general result in a decreased precision. We calculate the average recall and precision for the above experiments and plot their relation in Figure 4 that indicates a very good performance achieved by our system. At small number of returned shots the recall value was small while the precision value was very good. Increasing the number of returned clips increases the recall until it reaches one, at the same time the value of the precision was not degraded very much but the curve almost dwells at a precision value of 0.92. This way, the system provides a very good trade-off between recall and precision.

5.2. Experimental Results Using Unseen Queries

Here, the generalization ability of the system is tested by submitting queries that are not part of the video database. The experiments performed in the last section were repeated for three unseen clips. Figure 5 represents part of the retrieval interface displaying the first three hits obtained as a result of submitting an unseen query. Although, the system did not see the query shot before it searches its database for similar items and returns a number of shots with almost identical content (robot arms) as the query input. The generalization ability of our system is evident in this figure. In Figure 6, the relation between average recall and average precision was plotted and we can observe that the degradation in precision is reasonable given that the tested queries have not been seen before.

5.3. Similarity Model Performance

In this section, we analyze and evaluate the usefulness of the proposed similarity model. At first, the system performance will be evaluated using only one similarity factor at a time then the effect of using all of them simultaneously will be studied. To test the visual similarity factor, we repeated the experiments performed into the last two sections and got very good results. With respect to evaluating the frame rate similarity factor, the task becomes relatively easy for the system because of the straightforward way it can derive this information. Consequently, the obtained recall and precision results were very good. For example, when looking for shots that
have the same frame rate as shot 3 of the soccer clip, the system retrieves as hits the 61 shots in the database that has 24 f/s without any misses yielding ideal values for recall and precision. For the duration factor, we also obtained very good behavior of the system with ideal values for both recall and precision due to the same reason just mentioned. The system also achieves good performance when integrating these factors but results of these experiments are not listed here due to space constraints. Finally, we need to evaluate the time saving we gained by using the filtering stage to limit the scope of the search space. To accomplish this, a video clip will be submitted to the system as a query and the search time will be measured with and without filtering. Then, the time saving gained will be calculated. All other factors are kept the same and we use 41 clips in this experiment. The time saving averaged over all the used clips is 42%.

6. Conclusion

In this paper, we have presented our Video Content-based Retrieval system as a step to solve the general problem of accessing video data based on their contents. Moreover, a powerful tool to scope the size of the search space has also been introduced. Our main contribution in this research effort is the novel similarity model we proposed in which the similarity of video data is measured based on a number of factors that most probably reflect the way humans judge video similarity. The system enables users to select the weight associated with each of these factors in order to reflect the need of a particular user in a better way. The performance of the filtering stage has been evaluated and a considerable speed up gain of about 42% on average has been achieved. Evaluating the retrieval stage performance was achieved through measuring recall and precision of the returned results where very good values were obtained whether the query is part of the database or not.

References