Bilkent University at TRECVID 2004

S. Aksoy, K. Bircan, S. Çıracı, P. Duygulu, E. Karaca, S. Kasırga, T. Sevilmiş, M. Şener
Bilkent University
Department of Computer Engineering
Bilkent, 06800, Ankara, Turkey

Abstract

We describe our first-time participation, that includes three manual and interactive search runs, to the TRECVID video retrieval evaluation organized by the U.S. National Institute of Standards and Technology. We believe that this experience will benefit our future research and performance in the coming evaluations.

1 Introduction

This year, the Bilkent group participated in TRECVID for the first time. Our group (RETINA Vision and Learning Group) was recently established and TRECVID 2004 was the first project we were involved in. The group consists of two faculty members, three graduate students and three undergraduate students. Due to the insufficient resources and the steep learning curve regarding the data format and task specifications, we could submit only three runs to the search task. We take this year as a starting point for our future research on video classification and retrieval. The submitted runs were based on very simple ideas which do not represent our group's research directions. We believe that we will greatly benefit from the experience we have gained from TRECVID 2004 in our future research.

In all of our experiments, we have used the shot boundaries, keyframes and speech transcripts provided with the TRECVID 2004 data and the low-level features donated by the CMU group. The following sections briefly describe our approach for the runs we have submitted.

2 Text- and color-based search

Our system consists of keyword-based text search and color-based nearest neighbor search tools. The search process starts with a few keywords entered by the user according to the selected topic. A set of shots from the development data that include these keywords in their corresponding speech transcript are automatically selected and presented to the user to aid the visual query construction. The user selects a subset of these shots represented with their keyframes and submits this subset as the query. (We decided to include the shots from the development data to the query construction process because the shots supplied by NIST as examples for the query topics were not sufficient.) These keyframes submitted as the query are treated as positive examples in the rest of the search. An additional set of keyframes are randomly selected from the development data based on speech transcripts to form the negative example set. Then, a nearest neighbor [1] search is performed using RGB and HSV statistics as color features to find shots from the test data that are relevant to the query examples.

For the manual run, the set of shots obtained from the nearest neighbor search is intersected with another set obtained as a result of an automatic keyword search on the speech transcripts of the test data. The results of these intersections for each topic were submitted as the manual run. Because of the difficulties in finding useful development shots using the initial keyword search within the required time limits in the query construction, this manual run could retrieve shots from the test data only for a limited number of query topics.

For the interactive run, the user is allowed to select a new set of examples from the results of the initial search. These examples are incorporated into the nearest neighbor search with higher weights. This step is repeated within the allowed time limit and in each step some of the retrieved shots are eliminated based on a distance threshold. For most of the query topics, the number of iterations was limited to one because of the speed problems in the nearest neighbor search. In addition, only a few topics, such as tennis, golf, hockey, etc., could produce reasonable results in the interactive runs because of the system's strong dependency to color features.

3 Face-based search

For the queries related to finding a specific person such as Clinton, Netanyahu, Yeltsin, etc., we worked on an algorithm to learn the associations between faces and names. First, a text-based query is performed to obtain the shots including the name of the person in the speech transcript. Then, the faces in the keyframes are detected using the face detector provided by INRIA [2]. Since, this method produces many false positives when the threshold is set to a low value, a method based on finding skin color is applied on top of the face detector output in order to eliminate some of the false matches. The set of faces labeled as corresponding to a specific name includes the correct person, but also includes some other people such as anchors or reporters. In order to eliminate these false associations, the faces corresponding to a given name are clustered according to selected face features and then the dominant face cluster is associated with the given name. Features used in clustering are obtained by dividing each face area into blocks and then extracting the mean and standard deviation of the color values from these blocks. Unfortunately, search results based on this approach could not be completed before the submission deadline so we could not submit any related runs.

4 Conclusions

The runs we have submitted this year were based on simple algorithms that include keyword-based searches and nearest neighbor searches with color features. We believe that the experience we have gained from our first-time participation to TRECVID will benefit us greatly in our future research on video classification and retrieval.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley & Sons, Inc., 2000.

[2] K. Mikolajczyk. *Detection of local features invariant to affines transformations.* PhD thesis, INPG, Grenoble, France, 2002.