TREC Feature Extraction by Active Learning

J. Vendrig[1] J. den Hartog[2] D. van Leeuwen[3]
I. Patras[1] S. Raaijmakers[2] J. van Rest[2] C. Snoek[1]
M. Worring[1]

[1]Mediamill/ISIS, Informatics Institute, University of Amsterdam, Kruislaan 403,
1098 SJ Amsterdam, The Netherlands
[2]Mediamill/TNO-TPD, Netherlands Organisation for Applied Scientific Research,
Stieltjesweg 1, 2628 CK Delft, The Netherlands
[3]TNO-Human Factors, Netherlands Organisation for Applied Scientific Research,
Kampweg 5, 3769 ZG Soesterberg, The Netherlands

1 Introduction

Current multimedia retrieval research can be divided roughly into two camps. One camp is looking for the panacea which solves all problems in one system. The other camp focuses on very specific problems in restricted domains. In our opinion, the answer lies in the middle. A system should not desire to solve all problems, but should take advantage of a user's knowledge about his or her specific problem, so that the system can focus on it. On the other hand, available video analysis techniques should be extended to other domains which possibly were not envisioned upon their design. The challenge is transparent application of video analysis techniques to the appropriate user domains.

A user, especially an expert, has best knowledge about the characteristics of a particular domain. In this paper the user's input is given at index-time, rather than at query-time as done in our TREC 2001 contribution [2]. In [2], we associated user queries with video content descriptors via general Wordnet concepts. For example, query term *woman* maps to Wordnet hypernyms *"person, individual, human"* which we associated with the "face presence" descriptor. In this TREC 2002 contribution, we focus on building models for the association of content descriptors with generic concepts, such as the Wordnet hypernyms. Specifically, we focus on the ten generic concepts given by the TREC feature extraction task[1]. User and machine interact in order to map the semantic feature concept to content descriptors for a training set, so that shots can be classified for use in retrieval applications.

In this paper, we assume that every feature model is specific to not only the domain, but even to a collection, in order to exploit domain characteristics

[1] In the remainder of the paper we follow the TREC terminology, referring to the semantic concepts as features.

and user domain knowledge. The use of a small number of broadly applicable features for video content classification is described by IBM [11] in last year's TREC, showing relatively good results. Their probabilistic system mixes the use of models for general features (e.g. *outdoors* and *face*) with models for domain-specific features (e.g. *rocket* and *fire*). In our approach, every model is collection-specific. That is, even a general feature is considered to be specific. For example, although the *outdoors* semantic concept can be found in many video collections, its visual representation in a video may be quite different. Footage of the Discovery Channel shows a different kind of outdoors sceneries than television sitcoms. The old instructional videos for school kids in the TREC 2002 collection are quite different from the videos shown at school to the MTV- and Nintendo-generations. In addition, a feature can be defined differently amongst domains, resulting in the need for different models as well. For example, the definition of the *face* feature for a Cartoon Network collection is different from the one for C-SPAN.

Focusing on one particular collection enables for use, or some might say abuse, of simple content descriptors that are correlated to a semantic concept. This may be caused by the style of one or more people in the film crew, such as director, editor and camera man. An example of specific collection characteristics is found in the TREC 2001 video collection. The camera movement descriptor could be used to classify shots as mountains, as they roughly correspond with camera pans. Although such a classification method cannot be generalized to other collections, it allows for uncomplicated retrieval of videos in a specific collection.

We present a system which interactively learns user-defined semantic concepts for a specific collection from a domain expert. For each concept, the domain expert builds a model by feeding visual evidence to the system in the form of examples, without knowledge about the underlying classifier and descriptors. We employ a large set of multimedia descriptors for use in a Maximum Entropy classifier. The space for example selection is determined by the output of the incrementally improved model. The system is evaluated against the TREC 2002 feature extraction collection. The user information consists of the ten semantic concepts defined for the feature extraction task.

As our system is based on visual evidence, we focus on visual content of videos. That is, we focus on the features *outdoors, indoors, face, people, cityscape, landscape, text overlay* and *monologue.* The classification of the audio features (*speech* and *instrumental sound*) is provided independently and is described briefly in section 2.2.

The paper is organized as follows. In section 2 we describe how video content is represented for example selection and classification. In section 3 the use of the Maximum Entropy classifier for multimedia content is described. In section 4 the interactive selection of examples using active learning is explained. In section 5 we describe the experimental setup for TREC evaluation. Results are discussed in section 6. Finally, we present conclusions and future research in section 7.

2 Content representation

2.1 Data representation

The elementary unit in the context of TREC is a shot. However, the shot itself, i.e. the sequence of frames, is not always a good representation for the visual content. There are two reasons for employing an alternative representation. Firstly, a shot is not necessarily visually and semantically coherent. In [14] such fractions of a shot are called a shot-let, while in [12] this division is referred to as named events, which are short segments with a meaning that does not change in time. Division of the shot into smaller coherent fractions allows for better representation of the shot's content. In the context of TREC the further division is especially important, as the reference shot segmentation suffers from undersegmentation, combining consecutive shots into one shot. In addition, division into lower level units prevents loss of information due to aggregation. This is especially important in the context of TREC, as the feature definitions state that a shot is assigned to a class when at least an observable part of the shot belongs to that class. That is, a shot could be assigned to disjuncts concepts, e.g. both outdoors and indoor.

The second reason is computational feasibility. Using expensive descriptors derived from image processing on each frame in a shot requires a large amount of computing power. Meanwhile, the descriptor values can be expected to be highly similar in consecutive frames, as the content of frames change just gradually within a shot. Therefore it is expected that choosing a representation which requires less computing power does not result in loss of information as a consequence.

We choose to use content-dependent key frames as representation of a shot for image processing based descriptors. Motion descriptors are calculated on shot level for practical reasons, i.e. compatibility with existing systems. Content-dependent selection of key frames is based on the change in visual content during a shot compared to the change of content in the entire video [7]. Shots containing a relatively high amount of changes are assigned more key frames. Within a shot, key frames are chosen such that the total amount of change in the surrounding segment is approximately equal for all key frames in the shot.

2.2 Descriptors

Descriptors describe the content of a shot, or the content of a representation of a shot, in order to enable comparison of the content of two shots. For automatic use by the classifier, we employed a descriptor pool containing over 60 descriptors. The descriptor pool is not geared towards a specific data collection, as the system is designed to be independent of the data collection. The descriptors include atomic descriptors such as color values and color distributions, edge characteristics, and motion descriptors; and complex descriptors such as face presence [10] and camera movement [1]. Due to the large amount of descriptors, it cannot be assumed that they are independent. For example, one descriptor (e.g. dominant color) may coexist with a specialization of itself (e.g. dominant color in the top half).

In the following sections, we describe in more detail the descriptors that relate specifically to the temporal component of video shots.

2.2.1 Motion

Motion descriptors are extracted by a two-level analysis. At the block of frames level we analyze the motion by estimating a parametric motion model with a robust regression scheme [1]. In this way we obtain two types of low-level features. On the one hand we obtain the camera operation (pan, tilt, zoom-in, zoom-out or unknown) and the factors that are related to it (focus of expansion, pan-factor, etc). On the other hand, we obtain descriptors such as the average motion, the percentage of outliers from the dominant motion model, and the average position and motion of the outliers.

At the shot level, the descriptors of the block of frames level are combined for the estimation of descriptors such as the average pan factor in the shot, the motion activity due to camera operations and the percentage of frames in which the camera zooms in. The shot level descriptors are used for classification of the shot.

2.2.2 Audio

The classifications for the two audio features (speech and music) are provided independently by TNO-Human Factors. Classification is done without prior knowledge about the data set.

The speech/music discrimination is based on amplitude variation in the audio's signal envelope shape. Generally, speech has higher amplitude variations than music in the spectral regions around 475 and 2700 Hz. The input signal is partitioned into 1 second segments. For each segment, the amplitude fluctuations in these bands are determined and they are low pass filtered at 8 Hz. When the amplitude variation in either of the two spectral bands is above a certain threshold, the segment is identified as *speech*, otherwise as *music*. A third category *silence* is used if the total acoustic energy is low. The thresholds are based on values found for Dutch radio and television broadcast material, as well as eight music CD tracks with various music styles (classical instrumental, vocal, pop and jazz).

3 Classifier

The concept classifier used to model features has to deal with three issues concerning the data set and the descriptors. Firstly, the classifier has to cope with descriptors that are undefined for a shot. For example, when there is no pan camera movement in a shot, the pan factor descriptor is undefined. Secondly, in contrast with [11], it is our opinion that the classifier cannot assume descriptors are independent. Independence is not expected in multimedia objects, such as video shots, because they comprise several correlated information sources. In addition, the use of a large descriptor pool leads to interdependencies of descriptors. Thirdly, the classifier has to cope with an imbalanced data set. Just a relatively small number of positive examples is available in the training set. Training should focus on the positive examples, since the given features are one class problems [15].

For classification we choose the Maximum Entropy classifier, as it deals with the above issues. Firstly, it makes use of sparse vector format, thereby dealing with missing descriptor values. It is not necessary to provide dummy values for undefined descriptors. Secondly, the classifier assumes no independence between descriptors, in contrast to classifiers as Naive Bayes [8]. Thirdly, we found in experiments that a Maximum Entropy classifier is less sensitive to the majority effect than a Support Vector classifier [8]. Hence it is well suited for a data collection containing few positive examples.

Maximum Entropy classification has been applied successfully in a variety of domains, including the area of statistical natural language processing where it achieved state-of-the-art performance [4]. The Maximum Entropy framework was originally proposed by Jaynes [9] as a means to make inference based on partial information. Jaynes claimed that “the only unbiased assignment one could make, should use the probability distribution which has maximum entropy subject to whatever is known”, i.e. the probability distribution keeping the uncertainty maximal.

The Maximum Entropy approach allows for the use of a large amount of descriptors without the need to specify their relevance for training a specific semantic concept. The relative importance of each descriptor is computed automatically by the *Generalized Iterative Scaling* (GIS) algorithm [6]. This makes Maximum Entropy classification generally applicable.

A general problem for classifiers is diversity in descriptor types. The Maximum Entropy classifier suffers from this problem as well as it makes use of binary trigger descriptors, i.e. either the descriptor is true or it is false/undefined. Our original multimedia descriptors are both categorical and numerical. An example of the former descriptor type is “type of camera movement”, which takes values such as “zoom” and “pan”. Categorical descriptors can be used directly as a binary trigger. The mapping of numerical descriptor values to binary triggers, however, is non-trivial.

For further discretization of numerical descriptor values we employ a binning function which maps each value to a categorical representation. The binning process itself does not need to lead to loss of information. For many descriptors, numerical values have a higher precision than is needed or used. For example, there is no significant difference between value 0.10 and 0.101, i.e. more precision does not necessarily lead to better description of the content. It is often sufficient to express a descriptor’s value categorically. An example is the use of “mostly orange” or “very orange” to describe a color as done in [13]. The choice for the number of categorical values (bins) has to be established experimentally.

The binary trigger descriptor resulting from binning is used by the Maximum Entropy classifier. A disadvantage of binning which is specific to the Maximum

Entropy approach is the loss of order. That is, for the classifier bins are not related in any way. Bin 1 is not closer to bin 3 than to bin 9. Even when binning does not lead to loss of information for an individual descriptor value, it does lead to information loss for the similarity between shots.

The binning function has to take into account that not all descriptor values are normalized. Therefore we choose the use of equal frequency binning, which is performed after descriptor computation for the entire collection. It divides the descriptor values into a fixed number of bins such that each bin contains approximately the same number of values. Equal frequency binning is not sensitive to outlying values and skewed distribution of values over the range. The remaining problem is to choose the number of bins for discretization.

The implementation used for the Maximum Entropy classification is the publicly available OpenNLP Maximum Entropy Package [3].

4 Interactive teaching

As all learning classifiers, Maximum Entropy suffers from the need to select training examples from the collection. It requires a great deal of human effort to label the examples. Minimizing the human effort without compromising the quality of the examples is an important issue.

The traditional approach of random sampling is often chosen to acquire examples from a collection. However, for the problem at hand random sampling does not seem optimal. Cohn [5] addresses this issue in general in the context of neural networks, stating that in many formal problems it is more efficient to focus on a region of uncertainty rather than the entire collection. For the specific problem of TREC 2002 feature extraction we found this to be the case.

The reason to use more intelligent example selection than random sampling lies in the nature of the TREC features. Although the features can be perceived as binary (e.g. a shot contains *overlay text* or not), really the feature can be defined by positive examples only as the scope of negative examples is too broad. Therefore, focus on finding positive examples is needed.

In addition to the relative importance of positive examples, the positive examples are relatively rare in the collection. For example, we estimate 5% of the shots in the training collection contain *overlay text.* Even when a classifier does not need many positive examples, a sufficiently large set in absolute numbers must be given. Hence, it is more important to find positive examples describing the feature than to find examples representative for the collection.

To give precedence to labeling of positive examples we apply active learning, which is defined by [5] as “any form of learning in which the learning program has some control over the inputs it trains on”. That is, the classifier controls which examples are presented to the teacher for judgement.

Because of our focus on positive examples, for employment of active learning in our system we use a variant on the region of uncertainty described in [5]. Instead of presenting the teacher shots of which classification is most uncertain, we present shots for which labeling is most important, i.e. the shots most likely to be positive examples. This way, the teacher does not just provide examples, but indirectly he or she gives feedback on the classifier.

Theoretically, the active learning approach described could lead to undersampling of negative examples. However, in practice this would occur only when the model is very good from the start, or when positive examples are abundant in the collection. Both cases are unlikely in the context of the TREC feature extraction task.

Ideally, all presented examples would be classified by the teacher as either positive or negative. In practice, we have to introduce an "unclassified" category for two reasons. The first reason is that some shots do not contain sufficient information to be classified unambiguously. That is, the feature value cannot be determined without speculation or knowledge about the context. The second reason relates to definitions. The definitions given for the TREC experiment do not always match with the intuitive classification for a feature. This problem is not specific to TREC, but would occur in any large data collection for a broad domain, especially when there is more than one teacher. In cases where intuition and strict definition clash, the teacher uses the "unclassified" label. In terms of the classifier it means the model is not trained on "unclassified" examples, and we have no opinion on the outcome of the classifier for such examples when applied to the test collection.

We use the i-Notation system described in [16], extended with access to the OpenNLP Maximum Entropy Package [3] for selecting possibly positive examples. It is depicted in figure 1.

5 Experimental setup

In this section we describe briefly the most important parameters influencing the experiment.

Key frames are selected as described in section 2.1, with a minimum of 2 key frames per shot and an average (per video) of 1 key frame every 2 seconds.

The amount of bins is fixed to 4 for all numeric descriptors in the pool. Experiments with other bin amounts in the range of $3 \ldots 10$ showed no significant impact on results for the training set.

As the Maximum Entropy classifier's decision is binary for each feature, the results consists of shots for which a positive decision with a likelihood ≥ 0.5 is found.

The confidence of the existence of an audio feature (*speech* or *music*) within a video shot is computed as the number of segments classified as the feature normalized over the total shot length. The confidences of all shots are ranked and cut off at the given maximum of 1000 shots.

The two runs are different for the eight visual features only. The two aural features are constant. In the second run, we add the confidence measures for the aural features as descriptors used to classify the eight visual features.

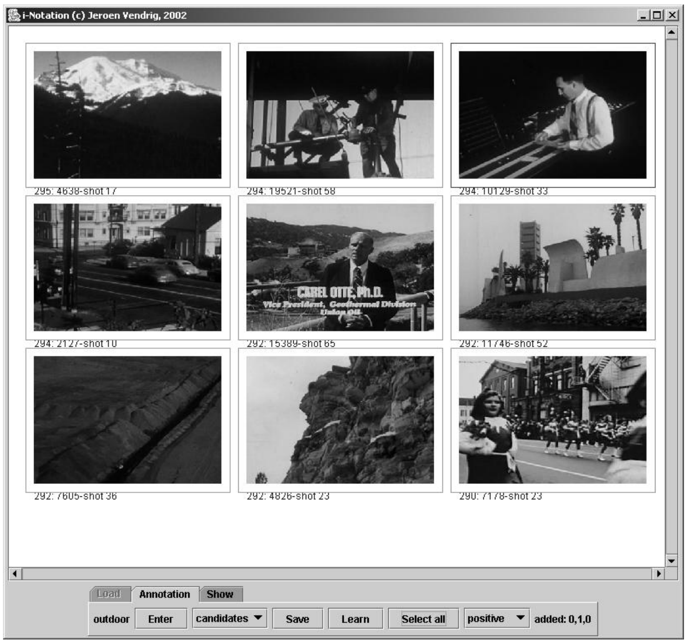

Figure 1: Screendump of the i-Notation system extended with active learning functionality.

6 Results and discussion

The likelihood threshold on the classifier decision confidence has a large impact on the evaluation, as the results lists are smaller than the maximum of 1000 shots for evaluation. The results lists vary from 744 shots for *outdoors* to 4 for *landscape*. The overall TREC results show that many more shots with the same feature are available in the collection, indicating that our threshold is too high.

Employment of a large descriptor pool does not have a negative impact on results according to a comparative experiment. We estimated results for the *face* feature using the "face presence" descriptor only, which is obviously a very powerful descriptor for this feature. Although adding the other descriptors do not lead to better classification results, they do not corrupt the classification either.

7 Conclusions and future research

The use of active learning in combination with the Maximum Entropy classifier leads to a generic approach for feature classification applying to a specific collection. The results of classification for the eight visual TREC features are not satisfactory. This may be due to the heterogeneity of the TREC collection, which is composed of several semantically unrelated collections. Although our approach should be able to cope with such a collection as well, the active learning component should be designed to take the heterogeneity into account. That is, the active learning component should take examples from all sub-collections for labeling by the user, thereby avoiding a local optimum.

The threshold used to determine positive decisions is too high. In the context of TREC evaluation, it would be better to rank all decision confidence values. Further research is needed to find out whether a ranking approach conflicts with the theory on which the Maximum Entropy classifier and its implementation are based.

The use of a large descriptor pool does not confuse the classifier when one very powerful, specific descriptor for a feature is present, as in the case of the *face* feature. Therefore, in future experiments we intend to use more descriptors rather than less. However, an automatic mechanism for selecting relevant descriptors from the pool for a particular feature is desired to be more robust against noise. Although initial experiments on the video data set do not yet show significant effects, we found selection and combination of descriptors to be useful in other Maximum Entropy applications.

The most important future research theme for use of Maximum Entropy classification in multimedia is the binning function. The effect of variations on the current binning function need to be measured. Examples of variations are small versus large amount of bins, determination of amount of bins for each descriptor individually, and using overlapping bins to deal with border values.

Acknowledgements

The authors wish to thank Jan Baan for providing camera work techniques and Thang Viet Pham for the face presence detection.

References

[1] J. Baan. Camera techniek detectie. Technical report, TNO, November 2000.

[2] J. Baan, A. van Ballegooij, J.-M. Geusebroek, D. Hiemstra, J. den Hartog, J. List, C. Snoek, I. Patras, S. Raaijmakers, L. Todoran, J. Vendrig, A. de Vries, T. Westerveld, and M. Worring. Lazy users and automatic video retrieval tools in (the) lowlands. In *Proceedings of the 10th Text Retrieval Conference (TREC)*, pages 104–113, Gaithersburg, USA, November 2001.

[3] J. Baldridge, T. Morton, and G. Bierner. The opennlp maximum entropy package. Technical report, SourceForge, 2002.

[4] A. Berger, S. Della Pietra, and V. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[5] D.A. Cohn, L. Atlas, and R.E. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[6] J.N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

[7] A. Hanjalic, G.C. Langelaar, P.M.B. van Roosmalen, J. Biemond, and R.L. Lagendijk. *Image and Video Databases: Restoration, Watermarking and Retrieval*, volume 8 of *Advances in Image Communication*. Elsevier, Amsterdam, 2000.

[8] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[9] E.T. Jaynes. Information theory and statistical mechanics. *The Phyiscal Review*, 106(4):620–630, 1957.

[10] T. V. Pham, M. Worring, and A. W. M. Smeulders. Face detection by aggregated bayesian network classifiers. *Pattern Recognition Letters*, 23(4):451–461, February 2002.

[11] J.R. Smith, S. Srinivasan, A. Amir, S. Basu, G. Iyengar, C.-Y. Lin, M.R. Naphade, D.B. Ponceleon, and B. Tseng. Integrating features, models, and semantics for trec video retrieval. In *Proceedings of the 10th Text Retrieval Conference (TREC)*, pages 96–103, Gaithersburg, USA, November 2001.

[12] C.G.M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, To appear.

[13] R.K. Srihari. Automatic indexing and content-based retrieval of captioned images. *IEEE Computer*, 28(9):49–56, 1995.

[14] H. Sundaram and S.-F. Chang. Determining computable scenes in films and their structures using audio visual memory models. In *Proceedings of the 8th ACM Multimedia Conference*, Los Angeles, CA, 2000.

[15] D.M.J. Tax. *One-class classification*. PhD thesis, TU Delft, 2001.

[16] J. Vendrig and M. Worring. Interactive adaptive movie annotation. *IEEE Multimedia*, To appear.