Video Frame Categorization Using Sort-Merge Feature Selection

Yan Liu and John R. Kender
Department of Computer Science
Columbia University
New York, NY 10027
{liuyan, jrk}@cs.columbia.edu

Abstract

Feature selection for video categorization is impractical with existing techniques. In this paper we present a novel algorithm to select a very small subset of image features. We reduce the cardinality of the input data by sorting the individual features by their effectiveness in categorization, and then merging pairwise these features into feature sets of cardinality two. Repeating this sort-merge process several times results in the learning of a small-cardinality, efficient, but highly accurate feature set. The cost of this wrapper method for learning the feature set, approximately O(F logF) where F is the number of incoming features, is very reasonable, particularly when compared with the impracticality of applying much higher cost current filter or wrapper learning models to the massive data of this domain. We provide empirical validation of this method, comparing it to both random and hand-selected feature sets of comparable small cardinality.

1. Introduction

The rapid growth and wide application of digital video data has led to a significant need for efficient video data management. Temporal video segmentation and classification is an important topic in video analysis. It is often the first data reduction step in any further processing, since videos, particularly those obtained after editing, appear to have a hierarchical structure, consisting of frames, shots, scenes, higher order segments, and the full video.

Koprinska and Carrato in [1] give an overview of existing techniques for video shot segmentation. These techniques operate on both uncompressed and compressed video streams, relying on definitions of visual similarity. However, in most situations, these algorithms do not reflect much semantic information of the video, which video database users are more concerned about. In response, Sundaram and Chang in [2] propose a more semantic scene segmentation algorithm, using visual and audio features together, and Zhou et al. propose in [3] a video classification method following a more supervised approach, although one that ends up being more application specific (in their case, basketball video event indexing).

At the same time, due to the huge volume of video data, the time complexity of segmentation and classification algorithms is another bottle-neck. Researchers therefore work on speeding up their algorithms; one way is by seeking efficient ways of reducing the dimensionality of the data prior to segmentation. Unfortunately, existing methods for feature selection can't be applied to this domain because of their time complexity.

This paper proposes a novel algorithm for frame categorization that follows the dimensionality reduction approach. It learns a small set of image features to use in classification, in a way that allows both high accuracy and high efficiency in frame categorization in video data. Section 2 introduces some related work of video scene categorization and feature selection. Section 3 proposes the sort-merge feature selection algorithm and provides the framework of scene categorization using this algorithm. Section 4 presents empirical validation of the accuracy of algorithm when applied to the particular genre of instructional videos. We close the paper with discussion and planned future work in section 5.

2. Related work

2.1. Scene categorization

As Huang et al. define in [4], scene categorization is the classification of a video sequence into one of a few predetermined scene types. Four different methods for integrated audio and visual information based on Hidden Markov Models are proposed in that paper; they classify TV programs into news reports, weather forecasts, commercials, basketball games, and football games. They note in their conclusion that reducing the input feature dimensions and choosing effective feature sets are the difficult problems, whose solution should lead to better performance.

Dimitrova et al. in [5] propose an algorithm for video classification using face and text "trajectories". Their work is based on the observation that in different TV categories there are different face and text trajectory patterns. This method relies heavily on the designers' intuitions about human perception and video editing grammars, and it is therefore difficult to extend these methods to other domains where the relation between the features and categories are unclear and changeable.

In contrast, Girgensohn and Foote in [6] describe techniques for classifying video frames into different categories using statistical models such as principal component analysis (PCA). This method makes use of the character of the data distribution in order to explore the relation between the features and the categories, including those that can not be readily perceived by a human experimenter. Additionally, it is readily extensible to use with uncompressed video. However, pure PCA is a data reduction technique of high computational cost, and its space requirements limit its ability to handle data sets on the order of the cardinality of the number of pixels in an image.

Approximation algorithms for dimensionality reduction, which have more reasonable time and space demands than pure PCA, have been proposed and demonstrated in the video domain, such the Fastmap method of Faloutsos and Lin [7]. Nevertheless, what results from such algorithms is a way to linearly map all incoming image features into as smaller set of derived features; it is still necessary to examine each feature of the image at classification time.

In this paper, we present a novel feature selection method which achieves simultaneously several goals: it categorizes frames of a video into semantically meaningful categories based on learned statistical regularities, operating on selected features rather than remapped ones. The selection operation has low time and space cost, and the subsequent use of the reduced feature set in classification is likewise highly efficient but highly accurate.

2.2. Feature selection

Feature selection methods have received significant attention in the artificial intelligence and learning literatures recently. They have been successfully used in classification applications with moderately large data sets, such as text categorization or genomic microarray data analysis. To our knowledge, however, they have not yet been applied to video scene categorization, quite possibly due to the massive amounts of computer time involved even in these more limited domains (on the order of weeks). Nevertheless, the categorization of video data shares several common characteristics with these existing domains. What this paper presents is a way to avoid the huge computational cost of existing feature selection methods.

A precise mathematical statement of the feature selection problem is not widely agreed upon, partly because there has been substantial independent work on feature selection in several fields: machine learning, pattern recognition, statistics, information theory, and the philosophy of science. Each area has formalized the definition from its own viewpoint, and each definition has been colored by the intended application.

However, there appears to be two major approaches. The first emphasizes the discovery of any relevant relationship between features and concept, whereas the second explicitly seeks a feature subset that minimizes prediction error. The first is referred to as a filter method, and it finds a feature subset independently of the actual induction algorithm that will use this subset for classification. Ordinarily, filter methods use simple statistics computed from the empirical feature distribution to select strongly relevant features and to filter out weakly relevant features before induction occurs; see Blum and Langley [8]. In contrast, a wrapper method searches the space of feature subsets, using cross-validation to compare the performance of a trained classifier on each tested subset, directly optimizing the induction algorithm that uses the subset for classification. As Xing et al. state in [9], wrapper methods attempt to optimize directly the predictor performance so that they can perform better than filter algorithms, but they require more computation time. Seen in this context, this paper proposes a novel sort-merge feature selection method with accuracy approaching that of a wrapper method but with a cost comparable to a filter method.

Feature selection methods are typically designed and evaluated with respect to the accuracy and cost of their three components: their search algorithm, their statistical relationship method (in the case of filter methods) or their induction algorithm (in the case of wrapper methods), and their evaluation metric (which is simply prediction error in the case of wrapper methods). The dominating cost of any method, however, is that of the search algorithm, since fundamentally feature selection is question of choosing one specific subset of features from the power set of features. This is an exponentially hard problem, and intractable if the set of features is very large as it is with image data. A more realistic design is to look for an approximate search algorithm that achieves high performance; this is necessarily a heuristic approach.

Three general kinds of heuristic search algorithms have been used: forward selection, backward elimination, and genetic. Forward selection starts with the empty set and successively adds individual features, usually following a variant of a greedy algorithm, terminating when no improvement is possible. However, it can't remove any features, and therefore ends up making what amounts to local optimizations to the growing set. Backward elimination, which does the reverse, starts with the full set of features and heuristically subtracts individual features. It suffers from a similar problem of local optimization, as removal of a feature is irrevocable. A genetic algorithm, which permits both the addition and deletion of features to a surviving population of evolving subsets of limited cardinality, is more likely to seek a global optimum. But it is computationally costly, and requires a more elaborate definition of algorithm convergence.

Current feature selection algorithms work well when there is a clear logical relationship between features and concept, such as a conjunction or disjunction of a list of functions produced by an induction algorithm. This "meaning" of the subset permits some insight for the experimenter's refinement of the design of the feature selection method. However, when the target concept has a more complex and unclear relation with features, or when the features tend to be related to each other, the design becomes more difficult. This is unfortunately the case with video frame data. The method we propose therefore attempts to finesse this problem of design by using the intrinsic redundancy of video data to the algorithm's advantage. The precise minimal subset cardinality is not sought directly; a heuristic approach to cardinality is combined with a disciplined genetic-like mixing of features to drive a selection process whose performance in classification is "good enough".

3. Feature selection for video

3.1. Sort-merge feature selection tree

Our sort-merge feature selection algorithm combines the features of forward selection, backward elimination, and genetic algorithms. To avoid irrevocable adding or subtracting, it always operates on some representation of the original feature space, so that at each step every feature has an opportunity to impact the selection. To avoid heuristic randomness, at each step a greedy algorithm is used to govern subset formation. Further, the recursive nature of our method provides an additional advantage over existing methods, in that it enables the straightforward creation of near-optimal feature subsets of any or all given cardinalities or accuracies, with little additional work. The entire sort-merge tree of progressively more accurate feature subsets can be stored, in space O(F log F), and accessed at a later time.

The sort-merge algorithm can be divided into two parts: the creation of a tree of feature subsets, and the manipulation of the tree to create a feature subset of desired cardinality or accuracy. Each part uses a heuristic greedy method.

Table 1 shows the algorithm of setting up the tree. Figure 1 illustrates a tree with F=256. Table 2 shows the algorithm of cutting the tree based on the application requirement, for example, to create a feature space with exactly r features.

```
Initialize level = 1
        F singleton feature subsets.
While level <  $\log_2$ F
        Induce on every feature subset.
        Sort subsets based on their
        classification accuracy.
        Combine, pairwise, feature subsets.
```

Table 1. Sort-merge feature selection basic algorithm.

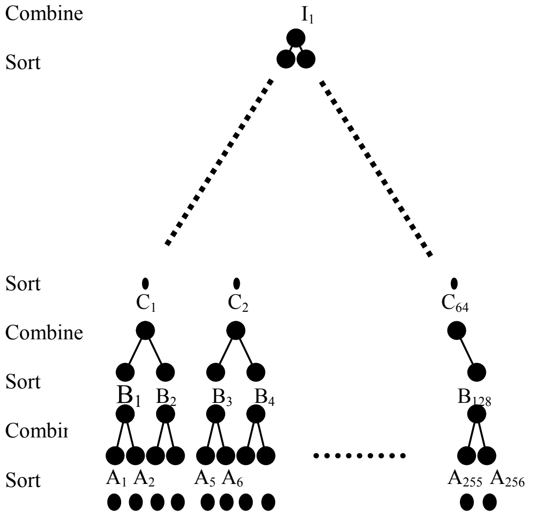

Figure 1. Setup of the sort-merge feature selection tree

The performance of a wrapper feature selection algorithm not only depends on the search method, but also on the induction algorithm. Like the search techniques, induction algorithms may have to be approximated for large datasets; some very common algorithms are infeasible for the cardinalities seen with video data.

```
Select the leftmost branch of size $2^{\lceil \log 2r \rceil}$.
Initialize cutout = $2^{\lceil \log 2\ r \rceil}$ - r.
While cutout >0
        Let branch-size = $2^{\lfloor \log 2\ cutout \rfloor}$.
        For all remaining branches of this
        size, evaluate the induction result of
        removing those branches individually.
        Remove the branch with best result.
        Let cutout = cutout – branch-size.
```

Table 2. Algorithm to select exactly r features from the tree of feature subsets.

For our induction method during the course of the learning, we use a novel combination of Fastmap for dimensionality reduction and Mahalanobis maximum likelihood for classification. We refer readers to the literature for a detailed explanation of these methods, but summarize their significance here.

In brief, Fastmap proposed in [7] approximates PCA, with only linear cost in the number of reduced dimensions sought, C, and in the number of features, F. It heuristically replaces the computation of the PCA eigenvector of greatest eigenvalue, which represents the direction in the full feature space that has maximum variation, with a (linear) search for the two data elements that are maximally separated in the space. The vector between these two elements is taken as a substitute for the eigenvector of greatest eigenvalue, and the full space is then projected onto the subspace orthogonal to this substitute vector. The process then repeats for the desired number of times. By the use of clever bookkeeping techniques, each additional dimension takes time approximately linear in the number of features. This linearity of the cost of Fastmap is a critical advantage, and permits its use for very large datasets.

In brief, as defined in statistical texts Duda et al. [10], or in the documentation of Matlab, the Mahalanobis distance computes the likelihood that a point belongs to a distribution that is modeled as a multidimensional Gaussian with arbitrary covariance. During training, each image in a training set for a video category is first mapped to a point in the space of reduced dimension C. Then the distribution of these mapped points is approximated by a C-dimensional Gaussian with a non-diagonal covariance matrix. Multiple categories and training sets are represented each with their own Gaussians. The classification of a test image is obtained by mapping it, too, into the reduced C-dimensional space, and then calculating the most likely distribution to which it belongs. That is, the classification label assigned to it is the label of the training set centroid to which it has the minimum Mahalanobis distance.

3.2. Framework of scene categorization

In this section, we develop the full sort-merge method and apply it to frame categorization of MPEG1 instructional videos. The method is a generic wrapper method for feature selection, and its principal contribution, that of a heuristic subset search technique for very large feature sets, is also applicable to other datasets of comparably large feature cardinality, to other video formats, and to other video contents and categories.

First, if it makes sense for the application, the video is down-sampled temporally, spatially, and/or spectrally. In our present experiments, we temporally subsample our MPEG1 video of a 75 minute-long lecture by using only every other I frame (that is, one I frame per second). We spatially subsample by using only using the DC terms of each macro-block of the I frame (consisting of six terms, one from each block: four luminance DC terms and two chrominance DC terms). Using DC terms as features to select from is very popular, see [1][12]. We do not spectrally subsample, but the method is transparent to this. This gives us, for each second of video, 300 macroblocks-worth (15 by 20) of 6 bytes (4 plus 2) of data: 1800 initial features. For purposes of learning, however, we consider the 6 DC terms to be an undecomposable vector, so our initial application more consists more accurately of 300 features per second of video, for about 4500 seconds. Each feature is placed into its own subset to initialize the sort-merge process; each feature subset has cardinality 1. In our application, we start with 300 such subsets, and have 4500 seconds of video to classify with these subsets.

Second, using Fastmap, the dimensionality of each feature subset is reduced to a pre-specified small number, C, of dimensions. (This makes more sense after the first several steps.) In our application, we ran experiments in which C varied from 1 to 10; a value of C of 2 or 3 was usually sufficient, however.

Third, for each feature subset at this level, using the reduced dimensionality representation, the training sets of the video train the induction algorithm to classify the test sets of the video. In our application, this meant that each training set was represented by a C-dimensional Gaussian, although other learning methods can be trained on the reduced representation. In our application, in the context of instructional video, we had four labels: instructor is writing an overhead slide, instructor is announcing, instructor is displaying a computer demo, and the class is discussing.

Fourth, the classification accuracy of each feature subset is measured. If any subset achieves the pre-specified desired accuracy, or if the cardinality of each subset achieves the pre-specified desired cardinality, the process stops, and that subset is the desired feature subset. Otherwise, the feature subsets are sorted by accuracy, and the next level of the feature subset hierarchy is formed by merging these subsets pair-wise and in order (see Figure 1). The cardinality of each feature subset doubles, but the number of such subsets is halved. Because of this, the amount of work at each level remains approximately constant, at O(F).

Fifth, the process repeats again, starting at the Fastmap step. It is clear that at most O(log F) iterations of this sort-merge algorithm are necessary.

4. Experiment

In our application, we have approximately 4500 seconds (units) of video to classify, 300 features for each unit, four classification categories, and about 400 units of training. Existing feature selection methods, which typically have been reported to run for several days on features sets of cardinality at least one decimal order of magnitude smaller, are intractable on this dataset; see Koller and Sahami [11]. Therefore, we compared the classification accuracy of our new method against two imperfect but feasible benchmarks, random feature selection, and hand feature selection: see the work of Xing et al who were similarly forced into such benchmarks [9]. These experiments used the same data and same induction methods; the only difference was how the feature subset was chosen. To simplify our presentation, only the comparison experiments that selected 30 features from the 300 are displayed here, although we do display the effect of varying the value of C.

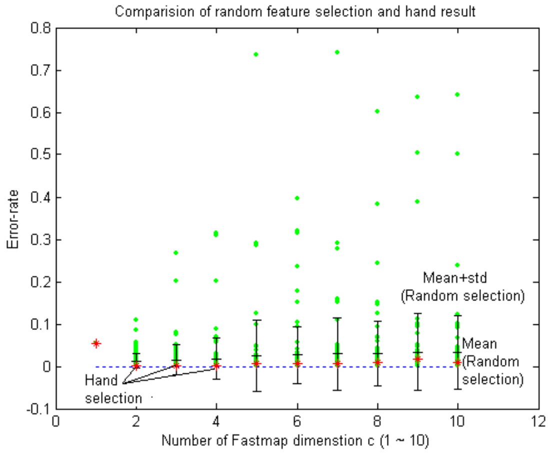

Figure 2. Classification results for random feature selection and hand feature selection (r=30).

For random feature selection, we ran 100 experiments in which 30 features were selected randomly. Points in Figure 2 show the error rate of scene categorization under different Fastmap reduced dimensions of C, from 1 to 10. The error bars are drawn at the mean error plus one standard deviation. The dashed-line shows the base of error rate of zero. Superimposed on the graph are asterisks representing the error rate of hand selection, typically about 0.2%. Figure 3 shows in more detail the error rate of scene categorization under the case of C=2, with each run of the random experiments illustrated. As expected, the rate of error is highly variable, with the standard deviation being larger than the mean.

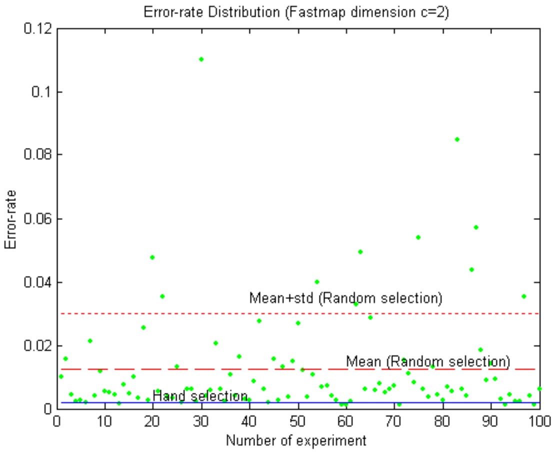

Figure 3. Classification results for C=2 (r=30).

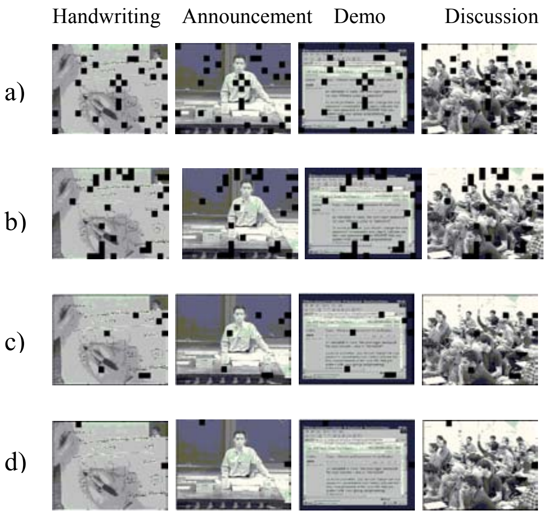

Figure 4. Select features using different methods.
(a) Feature selected by hand (r=30).
(b) Feature selected by sort-merge method (r=30).
(c) Feature selected by sort-merge method (r=8).
(d) Feature selected by sort-merge method (r=2).

Figure 4 (a) shows as black boxes those macro-blocks selected by hand, based on the intuition that the position of the instructor is important. Figure 4 (b) shows the ones selected by the sort-merge method; surprisingly, the method favors border macroblocks, with 20 of the 30 chosen at, or just one macro-block away from, the image border; this is possibly because these pixels tend to be the most stable over time. We note that this pattern of preferring stable background to dynamic foreground persists even when features are sparse. Figure 4 (c) shows r=8 features, only 3 of which are foreground, and Figure 4 (d) shows r=2 features, both background.

Figure 5 is a grand summary. The classification error rate of the sort-merge method is not only less than that of hand selection, but also appears to be very stable as the Fastmap dimension varies: this is critical, as C must be fixed before hand.

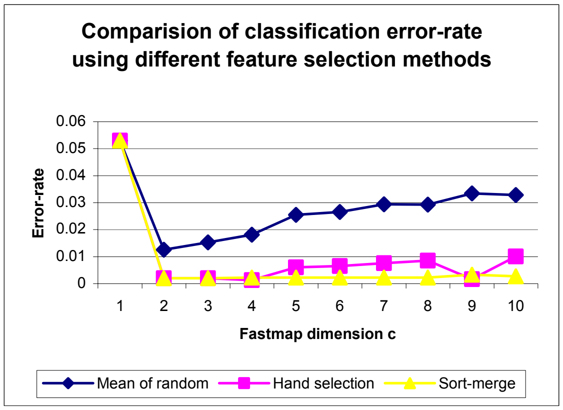

Figure 5. Scene classification result of different feature selection method (r=30).

5. Conclusion

We have presented a novel, low-cost, and accurate sort-merge method for selecting features for video frame classification, and have demonstrated some of its results on an extended video. Although not illustrated here, we have found that the classification by these highly reduced feature sets only fails within the centers of dissolves; further, these errors are entirely reasonable, consisting of picking the other category involved in the dissolve. (This is often a judgment call in the establishment of ground truth, which is as likely to be the fault of the human labeler). We intend to pursue this work theoretically, in proving some theorems about the limits of its near-optimality, and experimentally, by exercising it on different video genres to derive heuristics about the most appropriate way to set the value of C.

Acknowledgments:

This research was supported by NSF grant EIA-00-71954.

Reference:

[1] Irena Koprinska and Sergio Carrato, "Temporal video segmentation: A survey", *Signal processing: Image communication 16*, 2001, pp.477-500.

[2] Hari Sundaram and Shih-Fu Chang, "Video secne segmentation using video and audio features", *IEEE International Conference on Multimedia and Expo*, 2000.

[3] Wensheng Zhou, Asha Vellakial and C.-C. Jay Kuo, "Rule-based video classification system for basketball video indexing", *ACM Multimedia*, 2000.

[4] Jincheng Huang, Zhu Liu, Yao Wang, Yu Chen and Edward K. Wong, "Integration of multimodal features for video scene classification based on HMM ", *IEEE Signal Processing Society 1999 Workshop on Multimedia Signal Processing.*

[5] Nevenka Dimitrova, Lalitha Agnihotri and Gang Wei, "Video classification based on HMM using text and faces", *European Conference on Signal Processing*, 2000.

[6] Andreas Girgensohn and Jonathan Foote, "Video Classification using transform coefficients", *International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[7] Christons Faloutsos and king-Ip (David) Lin"FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets", *Proceedings of ACM SIGMOD,*1995, pp 163-174.

[8] Avrim L. Blum and Pat Langley, "Selection of Relevant Features and Examples in Machine learning", *Artificial Intelligence*, 97, pp.245-271.

[9] Eric P. Xing, Michael I. Jordan, Richard M. Karp, "Feature selection for high-dimensional genomic microarray data", *Proceedings of the Eighteenth International Conference on Machine Learning,* 2001.

[10] Richard O. Duda, Peter E. Hart and David G. Stork, *Pattern classification*, Wiley, New York, 2000.

[11] Koller, D. and Sahami,M. "Toward optimal feature selection", *Thirteenth International Conference on Machine Learning* , 1996.

[12] V. Kobla, D.S. Doermann and C. Faloutsos. "Video Trails: Representing and Visualizing Structure in Video Sequence", *ACM Multimedia Conference*, *1997,* pp. 335-346.