SELF-SUPERVISED LEARNING FOR ROBUST VIDEO INDEXING

Ralph Ewerth and Bernd Freisleben

SFB/FK 615 "Media Upheavals"
Department of Mathematics and Computer Science, University of Marburg, Germany
{ewerth, freisleb}@informatik.uni-marburg.de

ABSTRACT

The performance of video analysis and indexing algorithms strongly depends on the type, content and recording characteristics of the analyzed video. Current video indexing approaches often make use of thresholding techniques or supervised learning which requires labeling of possibly large training sets. Furthermore, the application of the same training model or parameters might lead to a sub-optimal indexing accuracy for a given video. In this paper, we propose to use a novel self-supervised learning framework for robust video indexing to address this issue. Based on an initial classification result for a given video, the best features are selected by Adaboost and are then used to train SVM (support vector machine) classifiers, all on the given video. Finally, a specialized ensemble of classifiers is employed for the given video for decision making. Experimental results show that a state-of-the-art video cut detection approach can be significantly improved by the self-supervised learning approach.

1. INTRODUCTION

The proliferation of digital videos is rapidly increasing in recent years, and thus the need for efficient retrieval techniques to support the search in large video databases is growing. Video indexing techniques generate meta information for video data that serves as a basis for search queries. Shot boundary detection, camera motion estimation, face detection/recognition, text detection/recognition, and topic detection are among the most important and popular indexing approaches. There is a great variety of video sources, video compression techniques and qualities, genres, which all together make correct video indexing difficult. Obviously, using the same approach or the same parameter settings will not be appropriate for each video. An analysis of proposed indexing approaches reveals that indexing performance varies and depends on the video content and other video characteristics. For example, a learned model might be too general for a video under consideration leading to a sub-optimal indexing performance. However, to the best of our knowledge, the issue of robustness has not been addressed explicitly yet.

In this paper, we propose a self-supervised approach to address the problem of robust video indexing. There are several contributions: Self-supervised learning is proposed for the first time in the context of video indexing. Furthermore, a novel self-supervised method is proposed that exploits both the properties of Adaboost and classifier ensembles using majority voting. Based on an initial classification result for a given video, the proposed approach utilizes Adaboost to find an optimal subset of features for this video. Then, this feature set is split into two complementary feature sets to train two SVMs directly on the given video. The splitting of the feature set is aimed at increasing the independence of the SVM classifiers. The independence is beneficial for the next step in which the baseline classifier and the SVMs are combined to form a specialized ensemble of classifiers for the given video using majority voting. Experimental results on the TRECVID 2005 test set show that the proposed approach improves the results of a high quality state-of-the-art video cut detection approach.

The paper is organized as follows. Related work is discussed in section 2. In section 3, the self-supervised learning approach is presented. The experimental settings and the results are described in section 4. Section 5 concludes the paper.

2. RELATED WORK

First, we review some state-of-the-art video indexing approaches (shot boundary detection, camera motion estimation and face detection) with respect to their robustness. Then, some self-supervised learning and co-training approaches are discussed.

Many proposals have been suggested in recent years for shot boundary detection. The TRECVID conference series is a forum that allows comparisons of different shot boundary detection approaches on the same test set with the same evaluation metrics. Smeaton and Over [8] show that submitted shot detection results vary in terms of recall and precision for different video sources ("NASA" and "News") used in the TRECVID 2005 evaluation (see Figure 1).

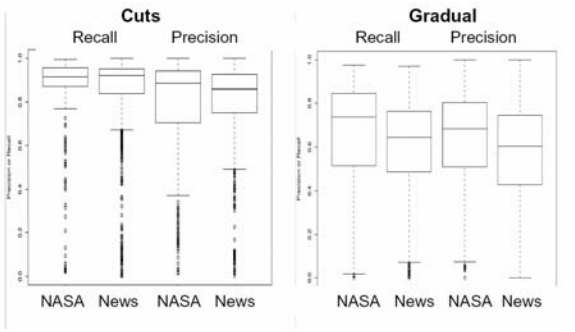

Figure 1: Distribution of recall/precision for shot detection depending of the video sources [8].

A closer look at successful video indexing approaches reveals that even top approaches are not designed to adapt to a particular video source. Typically, pre-defined thresholds or parameters are used, as exemplified by two of the best performing shot boundary detection approaches at TRECVID 2005:

Yuan et al. [13] combine a fade detector, a cut detector and a gradual transition detector for shot boundary detection. The fade detector is based on monochrome frame detection and tracking using several pre-defined thresholds. The authors apply a so-called graph partition model in which a graph is built based on pairwise frame similarities. One SVM is trained for cut detection, three SVMs are trained for gradual transition detection for different temporal resolutions using the TRECVID test sets of 2003 and 2004.

Tahaghoghi et al. [9] divide frames into 4*4 regions and disregard the frame center. Using a temporal sliding window, the pairwise similarity is computed for all frames using histograms of the HSV (Hue, Saturation, Value) color space. The sliding window is divided into pre-frames, the current frame and post-fames. The similarities are ranked with descending similarity and the ratio of pre-frames and post-frames in the top half represents the final similarity value. The sliding window size is a pre-defined parameter.

Considering the best TRECVID results 2005 for the task of camera motion estimation yields a similar picture. Yuan et al. [13] estimate the motion parameters of a two-dimensional affine model and finally apply thresholding rules to decide about the presence of motion. They achieved the best results for pan and tilt detection. Ewerth et al. [1, 3] obtained best results for zoom detection by computing the parameters of a 3D-camera model and applying thresholds for decision making.

Two of the most recent and successful face detection approaches employ machine learning and need a large training set. Schneiderman and Kanade [7] apply the wavelet transform and use wavelet features from various frequencies of different spatial resolutions to train a Naive Bayes classifier. Viola and Jones [10] train a cascade of Adaboost classifiers and mainly focus on real-time processing of video frames.

We have reviewed only some top performing approaches here, but to the best of our knowledge, self-supervised learning has not been applied in other video indexing approaches, too.

Up till now, there are only few applications of self-supervised learning or co-training in the field of pattern recognition. For example, Lieb et al. [5] propose a self-supervised approach for adaptive road following for driving vehicles to reduce the need that a road must be represented by unique identifying features. Wu and Huang [12] suggest self-supervised learning using labeled and unlabeled training data for object recognition in order to overcome the tedious and expensive task of labeling large training data sets. They extend a linear Discriminant-EM with a non-linear kernel. The experimental results show that their novel learning technique is competitive to SVMs and outperforms various approaches for hand-gesture recognition and fingertip tracking tasks. Oudot et al. [6] present a self-supervised method for writer adaptation in an online-text recognition system. In the self-supervised method, lexical results are compared with the classification hypothesis to find errors which are then used to re-estimate classifier parameters. Co-training (e.g. [11]) is a semi-supervised and multi-view learning approach which can be used if no sufficient amount of training data is available. The idea is to incorporate unlabeled data into the training and to make use of different feature sets (views) to train two classifiers. Wu et al. [11] suggest co-training for text detection in images. They train two SVMs on color respectively edge features and incorporate OCR into the training scheme.

3. SELF-SUPERVISED LEARNING FOR ROBUST VIDEO INDEXING

The key idea of the proposed approach is to use a robust baseline classifier for a given video X to automatically generate training data from the video itself. Then, a set of best features is selected for video X and split afterwards. The feature split is conducted to subsequently train different classifiers with a reasonable degree of independence on the video X using only the training data generated from the video itself. Kuncheva et al. [4] show that the independence of classifiers is advantageous to increase accuracy of an ensemble of classifiers. The system components (see Figure 2) are described in detail below.

First, a robust baseline indexing system, i.e. any classifier that proved to give stable results for the given task with a set A of features, is applied to obtain a first satisfactory result. Then, this result, including the classification errors, is used to select the best features for this video from a possibly large set B of features (where $A \subseteq B$, $|A| <= |B|$). Adaboost is applied (e.g. described in [10]) to obtain a ranking of the best features. These features are

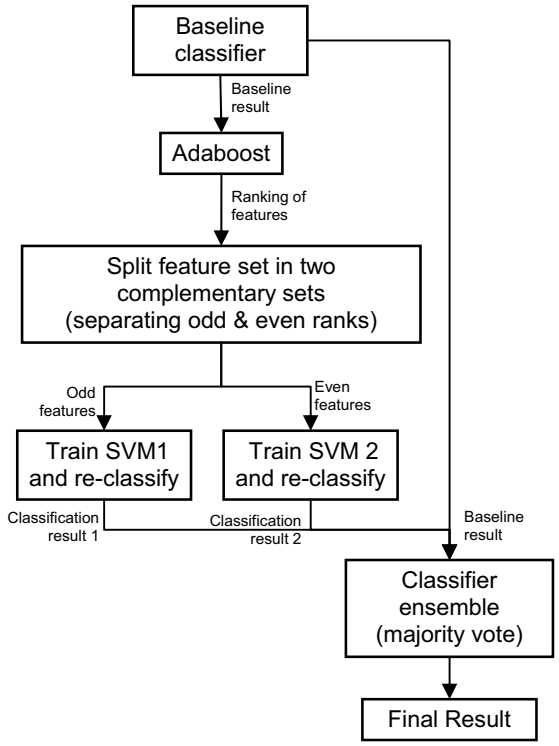

Figure 2: The prototype system of the self-supervised learning approach, applied to a given video X.

divided into two groups of odd and even features depending on the output ranking order computed by Adaboost.

The idea of the Adaboost approach is to combine a number of n "weak classifiers" to build a strong classifier within n rounds of training. For each feature, a minimum classification error is estimated. This classification error is computed based on the weights of the training samples that are weighted equally in the beginning. Misclassified training samples are re-weighted such that they have more impact in the next training round for the next "weak classifier". Thus, a selected feature has a higher probability to classify correctly those training samples that have been misclassified in preceding rounds. This property is the motivation to split the feature set depending on odd and even ranks for subsequent training. Then, one SVM is trained with the odd features and another SVM is trained with the even features directly on a given video X, again using the automatically labeled training data generated from the video X itself. Finally, the basic classifier and the two SVMs are combined to form an ensemble of classifiers using majority voting for the video X. There is evidence [4] that a reasonable degree of independence of ensemble classifiers improves accuracy respectively guarantees at least the accuracy of the weakest classifier in the ensemble if the classifiers' accuracy exceed a certain value.

The details of our prototype system of the learning framework for the task of video cut detection are as follows. Our previously proposed unsupervised clustering approach [2] is used as the baseline system. Only two features are used (motion compensated pixel differences, and the ratio of the second largest dissimilarity value divided by the local maximum within a sliding window of size 2m+1) in this approach in which an appropriate sliding window size is estimated automatically. For feature set B, we have defined 42 features for a certain frame distance describing frame dissimilarity with respect to:

- motion compensated pixel differences,
- histogram differences,
- luminance mean and variance,
- edge histograms of Sobel-filtered (vertically and horizontally) DC-frames,
- local histogram differences (3*3 regions), and
- ratio of second largest dissimilarity value divided by the local maximum for several sliding window sizes.

Two frame distances (1 and 2) are investigated resulting in a total feature number of 84. Then, n features are selected for a video using Adaboost. According to the ranking order, these features are split to train two SVMs. Together with the unsupervised system, they form an ensemble: a cut is detected if at least two of them vote that a frame is a cut.

4. EXPERIMENTAL RESULTS

The proposed self-supervised framework has been tested on the TRECVID 2005 shot boundary test set. The "MDC" library was used for MPEG decoding and the "libSVM" library for SVM implementation (Li and Sethi: iielab-ecs.secs.oakland.edu/demossoftware/MDC.html, and Chang and Lin: www.csie.ntu.edu.tw/~cjlin/libsvm/).

The proposed system (with n=11 features) has been implemented and compared with two of our submissions [3] to TRECVID 2005. The first submission consists of an unsupervised approach [2], the second submission consists of an ensemble of classifiers (Adaboost and SVM trained on a similar sets of features) in which each classifier has been trained on the TRECVID 2004 shot boundary test set. There are 2783 abrupt transitions in the video test set. Recall (R) is the number of correctly detected cuts divided by the total number of cuts, precision (P) is the number of correctly detected cuts divided by the total number of reported detections, including false alarms. The F1-measure is computed as follows: $F1 = 2*R*P/(R+P)$.

The experimental results are presented in Table 1. The precision is unusually low since there are many dissolves in two videos that have only one transitional frame and are often detected as cuts but are annotated as gradual transitions. The results show that the f1-measure increases from 0.854 to 0.878 using the self-supervised approach. The higher mean value of the f1-measures of the self-supervised approach is statistically significant and equals the f1-

measure of the supervised ensemble approach. For 8 out of 12 videos, the self-supervised approach leads to a lower total number of errors (including both false alarms and missed hits) than the supervised ensemble. The total number of errors decreases from 885 (baseline system) respectively 739 (supervised ensemble) down to 715 when the self-supervised approach is used. Thus, the approach can be recommended for cases in which costs for false alarms and missed hits are equal, gathering training data is difficult, or, in particular, in case of that the costs of false alarms are higher than those of missed hits. Overall, the self-supervised system is able to learn and automatically improve a model for a given video by itself without any pre-labeled training data.

5. CONCLUSIONS

In this paper, a novel self-supervised approach has been proposed to address the issue of robust video indexing. The approach is motivated by the analysis of video indexing approaches and issues related to the specifics and uniqueness of video source and content. Based on an initial classification result for a given video, the suggested approach utilizes Adaboost to select an optimal subset of features for this video. Then, this feature set is split into two complementary feature sets in order to train two SVMs on the given video. The splitting of the feature set allows to increase the independence of the classifiers, which is exploited in the subsequent execution of an ensemble of classifiers using majority voting. A prototype of the learning framework applied to video cut detection has been implemented and tested on the TRECVID 2005 test set. Experimental results indicate that the self-supervised system is indeed able to learn automatically by itself without any pre-labeled training data: The f1-measure is significantly higher than that of the baseline system and achieves similar detection results as an ensemble using supervised classifiers.

In the future, the impact of both classifier independence and baseline system accuracy will be investigated. Furthermore, other feature selection methods and classifiers will be analyzed. Finally, the application to other video indexing or pattern recognition tasks will be considered.

6. ACKNOWLEDGEMENT

This work is financially supported by the Deutsche Foschungsgemeinschaft (SFB/FK 615, MT project).

7. REFERENCES

[1] R. Ewerth, M. Schwalb, P. Tessmann, and B. Freisleben, "Estimation of Arbitrary Camera Motion in MPEG Videos", in Proc. of the 17th Int'l Conf. on Pattern Recognition, Vol. I, 2004, Cambridge, UK, pp. 512-515

[2] R. Ewerth and B. Freisleben, "Video Cut Detection without Thresholds", in Proc. of the 11th Int'l Workshop on Signals, Systems and Image Processing, 2004, Poznan, Poland, pp. 227-230

Measure	*K-Means Basis Approach*	*Supervised Ensemble Approach*	*Self-Supervised Approach*
F1	*0.854*	*0.878*	*0.878*
Precision	*79.0%*	*81.3%*	*83.7%*
Recall	*92.8%*	*95.4%*	*92.2%*
#Errors	*885*	*739*	*715*
#Misses	*200*	*129*	*216*
#False pos.	*685*	*610*	*499*

Table 1: Experimental results for the baseline approach, the ensemble including two supervised approaches, and the proposed self-supervised approach.

[3] R. Ewerth, C. Beringer, T. Kopp, M. Niebergall, T. Stadelmann, and B. Freisleben, „University of Marburg at TRECVID 2005: Shot Boundary Detection and Camera Motion Estimation Results", in Online Proceedings of TRECVID Conference Series 2005:
http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

[4] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. W. Duin, "Limits on the Majority Vote Accuracy in Classifier Fusion", in Pattern Analysis and Applications, 6, 2003, Springer-Verlag, pp. 22-31

[5] D. Lieb, A. Lookingbill, and S. Thrun, "Adaptive Road Following using Self-Supervised Learning and Reverse Optical Flow", in Online Proceedings of Robotics: Sciencs and Systems, Cambridge, USA, 2005:
http://www.roboticsproceedings.org/index.html

[6] L. Oudot, L. Prevost, A. Moises, and M. Milgram, "Self-Supervised Writer Adaption using Perceptive Concepts: Application to On-Line Text Recognition", in Proc. of the 17th Int'l Conf. on Pattern Recognition, 2004, Vol. II, Cambridge, UK, pp. 598-601

[7] H. Schneiderman and T. Kanade, "Object Detection Using the Statistics of Parts", in Int'l Journal of Computer Vision, 56 (3), 2004, pp. 151-177

[8] A. Smeaton and P. Over, "TRECVID 2005: Shot Boundary Detection Task Overview", in Online Proceedings of TRECVID Conference Series 2005:
http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

[9] S. M. M. Tahaghoghi, J. A. Thom, H. E. Williams, and T. Volkmer, "Video Cut Detection Using Frame Windows", in Proc. of the Twenty-Eighth Australasian Computer Science Conf., Vol. 38, Newcastle, Australia, 2005, pp. 193-199

[10] P. Viola and M. Jones, "Robust Real-Time Face Detection", in Int'l Journal of Computer Vision, Vol. 57 (2), 2004, pp. 137-154

[11] W. Wu, D. Chen, and J. Yang, "Integrating Co-Training and Recognition for Text Detection", in Proc. of IEEE Int'l Conf. on Multimedia and Expo 2005, Amsterdam, Netherlands, 2005, pp. 1166-1169

[12] Y. Wu and T. S. Huang, "Self-Supervised Learning for Visual Tracking and Recognition of Human Hand", in Proc. of the 17th National Conference on Artificial Intelligence, Austin, USA, 2000, pp. 243-248

[13] J. Yuan, L. Xiao, D. Wang, D. Ding, Y. Zuo, Z. Tong, X. Liu, S. Xu, W. Zheng, X. Li, Z. Si, J. Li, F. Lin, and B. Zhang, "Tsinghua University at TRECVID 2005", in Online Proceedings of TRECVID Conference Series 2005:
http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html