COMPARISON OF VISUAL FEATURES AND FUSION TECHNIQUES IN AUTOMATIC DETECTION OF CONCEPTS FROM NEWS VIDEO

Mika Rautiainen, Tapio Seppänen

MediaTeam Oulu, P.O.BOX 4500, FIN-90014 University of Oulu, Finland, tel. +358 8 553 2803
e-mail: mika.rautiainen@ee.oulu.fi, tapio.seppanen@ee.oulu.fi

ABSTRACT

This study describes experiments on automatic detection of semantic concepts, which are textual descriptions about the digital video content. The concepts can be further used in content-based categorization and access of digital video repositories. Temporal Gradient Correlograms, Temporal Color Correlograms and Motion Activity low-level features are extracted from the dynamic visual content of a video shot. Semantic concepts are detected with an expeditious method that is based on the selection of small positive example sets and computational low-level feature similarities between video shots. Detectors using several feature and fusion operator configurations are tested in 60-hour news video database from TRECVID 2003 benchmark. Results show that the feature fusion based on ranked lists gives better detection performance than fusion of normalized low-level feature spaces distances. Best performance was obtained by pre-validating the configurations of features and rank fusion operators. Results also show that minimum rank fusion of temporal color and structure provides comparable performance.

1. INTRODUCTION

Typically content-based video retrieval (CBVR) systems deal with low-level features that convey very little about the semantic content unless a trained system creates associations from the low-level features to a higher semantic context. For example, automatic detection of the presence of people support queries that attempt to locate a specific person from video database. Several studies have addressed the semantic feature, which is also described as semantic concept, detection [8][9][10][11].

In this study, semantic concepts mean textual terms that represent a conceptual entity in a video. They can be detected using automatic, semi-automatic or manual tools. Automatic detection holds a level of uncertainty whereas manual annotation can be subjective and laborious to create. An ensemble of classifiers can be trained for each concept. However, training of multiple classifiers for large concept lexicon can be tedious. A fast and simple method to build concept detectors was introduced in [6], where detectors were trained by selecting only small sets of positive examples for every concept.

This paper presents extended experiments with visual detectors for 12 semantic concepts from TRECVID 2003 semantic feature detection task [13]. The detectors use low-level visual features that measure video motion activity and spatial correlations of image gradients and colors. In comparison to prior research on visual detectors [13][6] this paper reports experiments with broader sets of concepts, low-level features, fusion techniques, training set sizes and larger test database. Section 2 describes selected low-level features and the fusion operations used in concept detectors. Section 3 describes the diverse experiments and Section 4 finalizes the paper with conclusions.

2. DETECTING SEMANTIC CONCEPTS

Semantic concept detectors create ordered video shot lists to describe the certainty of detection throughout the video database. In [6] we observed that several concepts coexist and correlate in a video, which is not suitable for multi-class classifiers. Our approach is to have several simplified concept detectors that are trained using small sets of positive example shots, each propagating labels to their nearest neighbors in selected feature spaces. The detection confidence is relative to the measurable low-level feature dissimilarity between the example and target.

2.1 Low-level Features

Features used in the detector measure motion, color and structure of a video shot. Dissimilarity between two feature vectors is measured using normalized city-block distance (L_1-norm). Short description of the used features follows:

Motion Activity (MA). MA is based on definitions of MPEG-7 Visual standard [4]. Following values describe the type of motion in the shot: discrete motion intensity; average intensity; short, medium and long runs of zero motion blocks.

Temporal Color Correlogram (TCC). TCC computes the autocorrelation of HSV pixel colors in the spatial neighborhood of the 20 temporally sampled video frames creating a vector of 432 feature values. Its effi-

0-7803-9332-5/05/$20.00 ©2005 IEEE

ciency against traditional color descriptors has been reported in [2][3]. TCC captures the probabilities for a pixel color to appear at given spatial pixel distances throughout a frame sequence. For a more detailed description of the algorithm, see [2].

Temporal Gradient Correlogram (TGC). TGC, initially used in the detector experiments in [6], describes spatial correlation of edge orientations in an autocorrelogram. The feature is computed from the 20 temporally sampled video frames in a shot. It depicts the dynamical compound of structural elements in a shot. Briefly, the Prewitt edges [12] are first detected and quantified from the sampled frames. Then the spatial autocorrelation is computed resulting a TGC vector of 20 feature values. More details about the algorithm can be found in [6].

2.2 Fusion of Low-level Features

Concept detectors are initialized with sets of K positive examples to produce result sets $R^f(k)$. The propagation of labels follows: First, dissimilarities to the example k in low-level feature space l results in rank-ordered list $D_l^f(k)$ where ks nearest neighbor has highest concept confidence. Subsequently $D_l^f(k)$ lists for every l 1…L are combined using either combination of ranks (Borda count variant) [5] or fuzzy Boolean combination of dissimilarity values:

$$r^f(k,n) = \quad (\frac{d_1^f(k,n)}{D_{1\max}^f(k)},...,\frac{d_L^f(k,n)}{D_{L\max}^f(k)}) \qquad (1)$$

$r^f(k,n)$ is overall rank or dissimilarity of a result shot n to the example k using L features. $d_l^f(k,n)$ is rank or dissimilarity to the example k of concept f in feature space l. $D_{l\max}^f(k)$ is largest rank or dissimilarity value to the query example k in its result set. is a set fusion operator: minimum rank (MIN), aggregation of ranks (SUM) or minimum dissimilarity (MINDIST)

2.3 Result Set Fusion

$R^f(k)$ contains a list of ranked database items and is a manifestation of confidence votes for a concept f based on overall similarity to the example k. Next, the ordered lists $R^f(1),..,R^f(K)$ are combined with a fusion operator to form a final confidence $s^f(n)$ for each item n. Finally, X top results are clipped for the evaluation procedure:

$$s^f(n) = \quad (\frac{r^f(1,n)}{R_{\max}^f(1)},...,\frac{r^f(K,n)}{R_{\max}^f(K)}) \qquad (2)$$

$$S^f = \lfloor sort\,\{s^f(1),...,s^f(N)\}\rfloor_X \qquad (3)$$

where $s^f(n)$ is the confidence of a shot n to contain concept f. $r^f(k,n)$ is rank or dissimilarity of item n to the query example k of concept f. $R_{\max}^f(k)$ is maximum rank or dissimilarity to the query example k in its result set. S^f is final ranked set of results for the concept f. $\lfloor\ \rfloor_X$ is X top-ranked items in a list. N is the size of the feature index.
 is a fusion operator: minimum-rank (MIN), rank-aggregation (SUM) or minimum distance (MINDIST).

3. EXPERIMENTS WITH NEWS VIDEO

The experiments were conducted in the framework of TRECVID 2003 semantic feature extraction task [1]. The task consisted of returning 2000 top ranked video clips for preset semantic features from a database of ~32000 MPEG-1 shot segments from ABC, CNN and C-SPAN. Following semantic concepts were used in this work: 'outdoors', 'people', 'building', 'road', 'vegetation', 'animal', 'car/truck/bus', 'aircraft', 'non studio setting', 'sporting event', 'weather news' and 'physical violence'.

3.1. Pre-validated Concept Detector Configurations

As a part of TRECVID 2003, IBM organized a joint collaborative video annotation effort to create a common ground truth for the development data [7]. 60 hours of shots were collaboratively labeled based on preset concept list and the annotations were distributed in MPEG-7 format. In this work the annotations were used in prior validation of the best detector configurations. The performance was compared against static configurations to find out the extent of gain for the computational validation.

First, positive example sets were selected from the development data as the input for the concept detectors. The sets were kept small. Sizes ranged from 7 for 'outdoors' to 26 for 'car/truck/bus'. Total count of positive examples was 217 and no negative examples were needed.

The validation was conducted by measuring the performance with different feature (MA,TCC,TGC) and rank fusion (MIN,SUM) configurations using the annotated truth data. The performance was measured from the detector output of 300 best ranked shots as the average of the precisions at correct detections.

Validation revealed two dominant detector configurations: First was a combination of TCC and TGC with , set to MIN (best performance in 'outdoors', 'road', 'animal', 'car/truck/bus'). Second used only color feature TCC with set to MIN (best performance in 'vegetation', 'sporting event', 'weather news', 'physical violence').

3.2. Semantic Concept Detection Experiments

For the actual test experiments detectors retrieved 2000 ranked shots from the test collection, which was not used during the development and validation phase. The evaluation used ground truth data created at NIST by pooling submitted results and creating relevance judgments [1].

The experiments evaluate the significance of low-level features, fusion operator configurations and pre-validation on detector performance. The effect of reduced training set was also tested with 106 examples (MT_extra3).

A single detection run consisted of detector outputs for all 12 features. Table 2 shows the overall run performance as the run-wise mean and median of the average precisions for all feature and fusion configurations. Average precision (AP) is a measure reflecting the performance over all relevant items in the result list, roughly depicting the surface under a precision recall curve [1]. The first row shows the performance of the validated configurations. Second row shows a fixed configuration of TCC and TGC features with , set to MIN. The first five runs (from MT1 to MT5) were used in the pooling of results at NIST whereas the runs from MT_extra1 to MT_extra5 were not contributing to it.

Table 1. Detector configurations and performance

RunID	Used Features			Mean	Med
MT1	*VALIDATED*	*VALID.*	*VALID.*	**9.0**	4.7
MT2	TCC/TGC	MIN	MIN	7.5	**5.6**
MT3	MA/TGC/TCC	SUM	MIN	3.7	2.5
MT4	MA/TGC/TCC	MIN	MIN	6.3	5.3
MT5	TGC	-	MIN	4.3	4.3
MT_extra1	TCC	-	MIN	7.8	3.3
MT_extra2	TCC/TGC	SUM	MIN	6.6	3.0
MT_extra3	TCC/TGC	MIN	MIN(106)	5.7	4.5
MT_extra4	TCC/TGC	MIN	SUM	3.9	1.8
MT_extra5	TCC/TGC	MDIST	MDIST	3.6	3.2

Run-wise means of the average precisions show that the validated detector configurations in MT1 obtained the best overall detection. Best median was obtained with fixed detector configuration of TGC and TCC together with MIN operators (MT2).

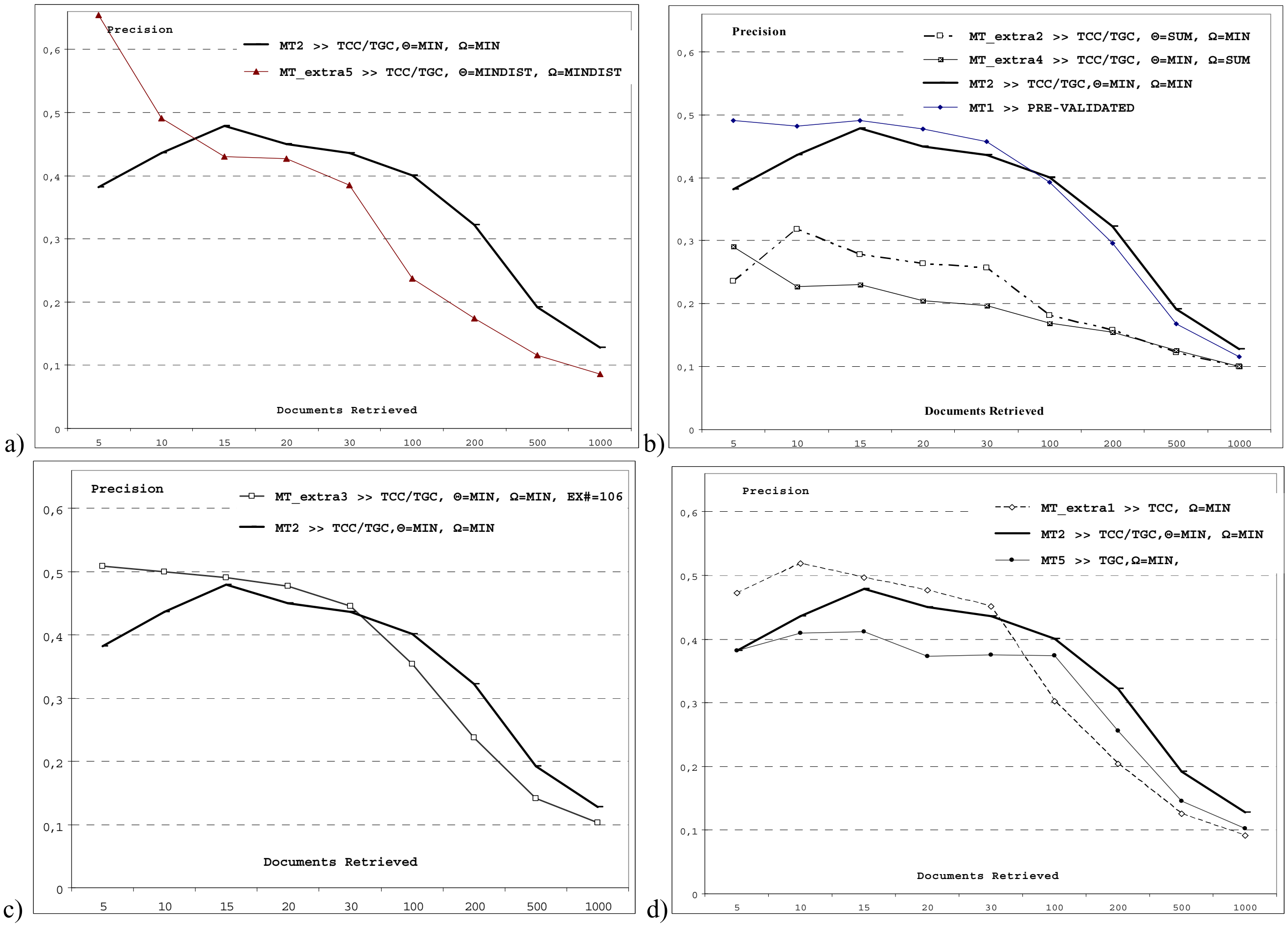

Figure 1. Detector precisions at nr. of shots retrieved (excluding 'weather news'). **a)** rank vs. metric based fusion operators **b)** the effect of various fusion operators **c)** small vs. large example sets **d)** color vs. structure

Using motion activity decreased multi-feature detector performance. As for individual performance TCC run defeated TGC run by performing better in ‘weather news’, ’vegetation’, ’aircraft’, ‘sporting event’, and ‘physical violence’. Overall best detection performances were obtained in ‘weather news’ (run MT1: AP 0.501), ‘sporting event’ (MT_extra4: AP 0.152) and ‘people’ (MT1: AP 0.096).

Figure 1 shows the detector precisions for selected runs without the performance of ‘weather news’, since the structure of weather news in the test collection favors the color feature TCC. At the top left figure, rank based MIN operator outdoes MINDIST, which selects the minimum of normalized L_1 distances to combine features and examples. MINDIST is weaker although it preserves the dynamical structure of the feature space unlike rank fusion. Figure at the top right corner shows that MT2 with MIN operator outperforms runs with SUM (aggregated ranks) and is nearing the performance of the validated run (MT1). Small (106) and large (217) example sets are contrasted at the lower left figure. With fewer examples initial precision is increased with the cost of recall. Figure at the lower right plots the difference between TCC and TGC performance. The steeper curve of TCC indicates that the local color correlation is effective only when the structure of color is playing a vital role in the context of a semantic concept, as in ‘vegetation’, otherwise the feature becomes confused. TGC feature is based on intensity gradients and is less restricted to specific visual settings.

4. CONCLUSIONS

Semantic concept detection experiments in a large news video database were presented in this work. Training is simple and computationally inexpensive. Unlike traditional classifiers, only small positive example sets are required in training.

Experiments showed that the rank-based fusion of features results in better overall performance than the fusion based on normalized low-level feature vector distances. Also minimum rank fusion is more effective than aggregation of ranks. Increasing the example set size was found to improve recall but degrade initial precision. TCC is particularly effective in concepts where color dominates the visual context, but is limited into fixed chromatic settings. Combination of TCC and TGC provides a good trade-off. In the future, detection performance could be improved using weights. Also performance against traditional classifiers should be inspected.

ACKNOWLEDGEMENTS

We would like to thank the National Technology Agency of Finland (Tekes), Academy of Finland and Nokia Foundation for supporting this research.

11. REFERENCES

[1] TREC Video Retrieval Evaluation. http://www-nlpir.nist.gov/projects/trecvid/ (4.1.2005)

[2] M. Rautiainen, and D. Doermann, “Temporal color correlograms for video retrieval.,” *Proceedings of 16th International Conference on Pattern Recognition*, Quebec City, Canada, 2002.

[3] J. Huang, S.R. Kumar, M. Mitra, and W.J. Zhu, “Image indexing using color correlograms,” *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp. 762-768, 1997.

[4] B.S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Language*, Wiley, John & Sons, Inc., 2002.

[5] T. Ho, J. Hull, and S. Srihari, “Decision combination in multiple classifier systems,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 16(1) , pp. 66–75, 1994.

[6] M. Rautiainen, T. Seppänen, J. Penttilä, and J. Peltola, “Detecting semantic concepts from video using temporal gradients and audio classification,” *International Conference on Image and Video Retrieval*, Urbana, IL, pp. 260-270, 2003.

[7] C.-Y. Lin, B.L. Tseng, and J.R. Smith, “Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets,” *NIST TREC-2003 Video Retrieval Evaluation Conference*, Gaithersburg, MD, November 2003.

[8] M.R. Naphade, T. Kristjansson, B. Frey, and T.S. Huang, “Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval in multimedia systems,” *In proceedings of International Conference on Image Processing*, vol. 3, pp. 536 -540, 1998.

[9] N. Haering, R.J Qian, and M.I. Sezan, “A Semantic Event Detection Approach and Its Application to Detecting Hunts in Wildlife Video,” *IEEE Transactions on Circuits and Systems for Video Technology,* Vol. 10(6)*,* pp. 857 – 868, 2000.

[10] S.F. Chang, W. Chen, and H. Sundaram, “Semantic visual templates – linking features to semantics,” *In Proceedings of IEEE International Conference on Image Processing*, vol. 3., pp. 531-535, 1998.

[11] A. Del Bimbo, “Expressive semantics for automatic annotation and retrieval of video streams,” *IEEE International Conference on Multimedia and Expo*, Vol.2. pp. 671-674, 2000.

[12] J.M.S. Prewitt, “Object enhancement and extraction,” In B.S.Lipkin and A. Rosenfeld, (eds) *Picture Processing and Psychopictorics*, Academic Press, New York, 1970.

[13] M. Rautiainen, J. Penttilä, P. Pietarila, K. Noponen, M. Hosio, T. Koskela, S.M. Mäkelä, J. Peltola, J. Liu, T. Ojala, and T. Seppänen, ”TRECVID 2003 experiments at MediaTeam Oulu and VTT,” *TRECVID Workshop at Text Retrieval Conference TREC-2003*, Gaithersburg, MD, 2003.