

TRECVID 2007 High Level Feature Extraction experiments at JOANNEUM RESEARCH

Roland Mörzinger and Georg Thallinger
JOANNEUM RESEARCH, Institute of Information Systems and Information Management
Steyrergasse 17, 8010 Graz, Austria
roland.moerzinger@joanneum.at

ABSTRACT

This paper describes our experiments for the high level feature extraction task in TRECVID 2007. We submitted the following five runs:

- A_jr1_1: Baseline run using early fusion of all input features.
- A_jr1_2: Classic early feature fusion and concept correlation.
- A_jr1_3: Classic late feature fusion.
- A_jr1_4: Late feature fusion and concept correlation.
- A_jr1_5: Early fusion of heuristically defined feature combinations.

The experiments were designed to study both, the performance of various content-based features in connection with classic early and late feature fusion, the influence of manually (heuristically) selecting input feature combinations and the application of concept correlation.

Our submission made use of support vector machines based on a variety of image and video features. The results of the experiments show that four out of five runs achieved a performance above the TRECVID median, including a run with 18 out of 20 evaluated high level features equal or above the median compared with inferred average precision. The mean inferred average precision of our baseline run is 0.056. Early fusion performed slightly better than late fusion on average, although the latter produced more scores above the TRECVID median. The experiment on concept correlation generally impaired the performance and outscored the baseline only for a few features. Heuristic low-level feature combinations displayed a rather poor performance. We assume that the good baseline is due to the effective grounding of a variety of low-level visual features and the generalization capability of the SVM framework with high-dimensional feature spaces.

1. INTRODUCTION AND OVERVIEW

In TRECVID 2007 our group participated in the BBC Rushes Summarization [2] and for the first time independently in the High Level Feature Extraction task. This notebook paper describes the submission to the High Level Feature Extraction task.

Taking part for the first time in this task, our main aim was to build a machine learning machinery in an integrated framework capable of processing the huge amount of data. A set of 5 runs were submitted with the goal to investigate:

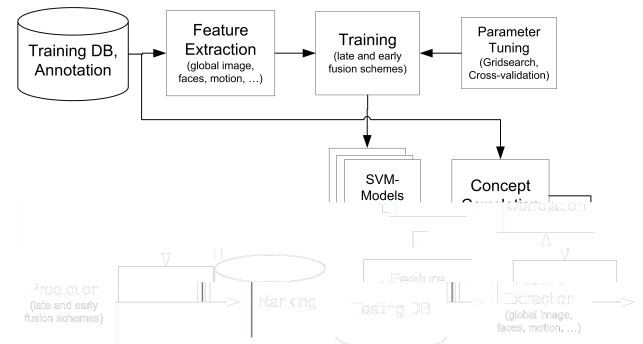


Figure 1: Basic system architecture.

- the suitability of various content-based features for building a baseline run.
- performance of early fusion as opposed to late feature fusion [7].
- the influence of application of simple a-priori knowledge about the correlation of concepts on the results.
- the influence of manual (heuristic) selection of low-level feature combinations on the results.

Figure 1 shows the basic system architecture. The annotated training data was obtained from the TRECVID 2007 collaborative annotation project [1]. For the implementation of the training and prediction components we use the LIBSVM software package [4]. The rest of this paper is organized as follows: Section 2 describes the used content-based features in detail, Section 3 and 4 outline the training and prediction process. Results are presented in Section 5.

2. CONTENT-BASED FEATURES

The content-based features described here are the basis for the detection methods for high-level semantic concepts. The following MPEG-7 [6] image features were extracted globally:

Color Layout describes the spatial distribution of colors. This feature is computed by clustering the image into 8x8 blocks and deriving the average value for each block. After computation of DCT and encoding, a set of low frequency DCT components is selected (6 for the Y, 3 for the Cb and Cr plane).

Dominant Color consists of a small number of representative colors, the fraction of the image represented by each

color cluster and its variance. We use three dominant colors extracted by mean shift color clustering [5].

Color Structure captures both, color content and information about the spatial arrangement of the colors. Specifically, we compute a 32-bin histogram that counts the number of times a color is present in an 8x8 windowed neighbourhood, as this window progresses over the image rows and columns.

EdgeHistogram represents the spatial distribution of five types of edges, namely four directional edges and one non-directional edge. We use a global histogram generated directly from the local edge histograms of 4x4 sub-images.

Gabor Energy is computed by filtering the image with a bank of orientation and scale sensitive filters and calculating the mean and standard deviation of the filtered outputs in the frequency space. We applied a fast recursive gabor filtering [8] for 4 scales and 6 orientations.

The extraction algorithm for *camera motion* is the same we used for the TRECVID 2005 camera motion task [3]. It is based on feature tracking using the Lucas-Kanade tracker, which is a compromise between spatially detailed motion description and performance. The feature trajectories are then clustered by similarity in terms of a motion model. The clustering algorithm is an iterative approach of estimating a motion parameter sequence for a set of trajectories and the re-assigning trajectories to the best matching parameter sequence. The cluster representing the global motion is selected. The decision is based on the size of the cluster and its temporal stability. According to the parameter sequence representing the dominant motion, the presence of pan, zoom and tilt is detected. For one or more segments per shot the following types of motion are described: pan left/right, tilt up/down, zoom in/out and static.

The *visual activity* feature is computed by temporally subsampling the video and computing the mean absolute frame differences (MAFD). The description contains statistics about minimum, maximum, mean and median MAFD per shot.

For each shot the *number of faces* is detected on the temporally subsampled video, by using the face detection method implemented in OpenCV ¹. The mode (most frequent value) of the number of detected faces in the frames is described for each shot.

2.1 Feature Preprocessing

The used content-based features capture both, global image properties (color and texture) and shot properties (faces, visual activity and motion). To overcome the limitations of having only one keyframe representing a shot's visual content, we extracted multiple frames per shot. For each of the extracted frames, a training or testing sample was created.

Further, some input features were preprocessed before transferring them to the training or prediction system. Specifically, the *number of faces* value was quantized to 0 (no face), 1 (one face) and 2 (two or more faces). This seems valid and necessary considering e.g. the high level features 'Crowd', 'Meeting' and 'Face', where the exact number of detected faces (e.g. 5 vs. 7 faces) is insignificant for the concept detector. Similarly, the *camera motion* is reduced from detailed pan, zoom and tilt information to 0 (no camera motion) and 1 (camera motion exists). As feature for *visual activity* the mean statistics was adopted. Finally, all

feature vectors were statistically normalized, by converting into a normal distribution with zero mean and unit variance.

3. TRAINING

Our approach to high level feature extraction is based on training support vector machines (SVMs) since they had achieved quite satisfactory performance in concept detection over the past few years. The classification of each high level feature (concept) was regarded as a two-class problem, where the positive and negative examples were extracted from the TRECVID 2007 Collaborative Annotation [1]. Since there were more negative examples than positive examples for most of the concepts, the SVM training data was composed of all positive annotations with a comparable number of randomly selected negative annotations. For better comparability we assured that the random selection produced the same annotations across different runs.

We adopted the Gaussian RBF kernel function. For each SVM a grid search was performed with cross-validation to select the best choice of the parameters C and γ .

Depending on the setup of our runs, SVMs were built using a single modality of low-level features (for late feature fusion) or using an input feature vector composed of multiple features. For one run we manually determined combinations of low-level input features for training, based on heuristics. For example, the low-level feature combination for 'Sky' was set to all MPEG-7 image features; for 'Sports' the low-level feature camera motion, visual activity and faces were also included.

4. PREDICTION

For the runs with early feature fusion we ranked the shots according to the outputs from the two-class SVMs.

Late fusion requires the combination of the output of various base classifiers. The first approach was to set up basic fusion operators such as the minimum, maximum, average and product of probabilities. For the experiment we empirically decided on an ensemble classifier that makes an overall prediction based on the product of probabilities. Of course, it would be preferable to learn an effective combination method, e.g. by means of a high level SVM or AdaBoost. However, this was not implemented due to time constraints. For some runs the prediction scores were modified by application of concept correlation, as described below.

4.1 Concept Correlation

The TRECVID 2007 collaborative annotation effort gives information about the annotations with high level concepts for all shots from the training set. From that we can compute a concept correlation matrix, i.e. a 36x36 matrix where for each concept the number of co-occurrences with other concepts is given. Figure 4.1 visualizes a normalized correlation matrix of 36 concepts sorted by the number of correlations. This plot reveals e.g. that the concept 'Computer.TV-screen' correlates with 'Person', 'Face' and 'Office'. The idea of concept correlation is to use this information to correct the confidence scores obtained from the prediction step. For example, an originally low confidence score for the concept 'Office' could be increased in the presence of high scores for 'Computer.TV-screen' and 'Face'. For that purpose the confidence values for the 36 concepts are cor-

¹<http://sourceforge.net/projects/opencvlibrary>

Feature	median	A_jr1_1	A_jr2_2	A_jr3_3	A_jr4_4	A_jr5_5
1: Sports (282)	0.028	0.023	0.006	0.014	0.014	0.019
3: Weather (34)	0.002	0.005	0.001	0.005	0.005	0.004
5: Office (1131)	0.061	0.083	0.072	0.061	0.042	0.026
6: Meeting (742)	0.053	0.101	0.120	0.065	0.033	0.051
10: Desert (67)	0.008	0.022	0.017	0.049	0.050	0.007
12: Mountain (124)	0.030	0.028	0.052	0.032	0.032	0.023
17: Waterscape_waterfront (902)	0.167	0.170	0.163	0.196	0.194	0.051
23: Police_Security (255)	0.003	0.004	0.004	0.018	0.010	0.002
24: Military (430)	0.005	0.034	0.007	0.005	0.004	0.004
26: Animal (845)	0.074	0.102	0.064	0.082	0.082	0.049
27: Computer_TV-screen (529)	0.051	0.103	0.026	0.135	0.114	0.108
28: Flag-US (12)	0.000	0.000	0.000	0.000	0.000	0.000
29: Airplane (58)	0.022	0.043	0.036	0.057	0.057	0.004
30: Car (672)	0.082	0.139	0.121	0.110	0.110	0.023
32: Truck (126)	0.026	0.055	0.053	0.046	0.046	0.042
33: Boat_Ship (297)	0.083	0.118	0.091	0.083	0.083	0.029
35: People-Marching (270)	0.028	0.046	0.054	0.050	0.050	0.007
36: Explosion_Fire (46)	0.005	0.009	0.029	0.005	0.005	0.002
38: Maps (117)	0.035	0.020	0.001	0.035	0.033	0.018
39: Charts (115)	0.017	0.019	0.006	0.004	0.004	0.001
mean IAP	0.039	0.056	0.046	0.053	0.048	0.024
Nr. $\geq$ median		17	13	18	14	4

Table 1: High Level Feature Extraction Results. The 20 evaluated features are shown with the number of positive training samples, the TRECVID median and the inferred average precision measures of all our runs. The mean Inferred Average Precision and the number of features above the median is stated.