The Influence of Cross-Validation on Video Classification Performance

Jan C. van Gemert, Cees G.M. Snoek, Cor J. Veenman and Arnold W.M. Smeulders
ISLA, Informatics Institute University of Amsterdam
Kruislaan 403, 1098 SJ
Amsterdam, The Netherlands
{jvgemert, cgmsnoek, cveenman, smeulders}@science.uva.nl

ABSTRACT

Digital video is sequential in nature. When video data is used in a semantic concept classification task, the episodes are usually summarized with shots. The shots are annotated as containing, or not containing, a certain concept resulting in a labeled dataset. These labeled shots can subsequently be used by supervised learning methods (classifiers) where they are trained to predict the absence or presence of the concept in unseen shots and episodes. The performance of such automatic classification systems is usually estimated with cross-validation. By taking random samples from the dataset for training and testing as such, part of the shots from an episode are in the training set and another part from the same episode is in the test set. Accordingly, data dependence between training and test set is introduced, resulting in too optimistic performance estimates. In this paper, we experimentally show this bias, and propose how this bias can be prevented using *episode-constrained* cross-validation. Moreover, we show that a 17% higher classifier performance can be achieved by using episode constrained cross-validation for classifier parameter tuning.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*

General Terms

Experimentation, Performance, Reliability

Keywords

Multimedia performance evaluation, cross-validation, parameter tuning, semantic concept detection

1. INTRODUCTION

Machine learning techniques have proven to be a valuable addition to the repertoire of a multimedia researcher. Applications of machine learning techniques in multimedia are found in semantic video labeling [10], video shot detection [9], audio classification [5], scene recognition [12], sports analysis [2], emotion recognition [1], Meeting analyse [6], and in many other areas. Moreover, multimedia researchers have contributed to specifically designed classifiers for multimedia analysis [7].

A prerequisite to good classification results is an accurate estimation of classifier performance [3, 4]. The estimated classification performance may be used in finding the best parameters of the classifier and helps deciding between different features. Hence, the estimated classification performance determines the quality of the classification results.

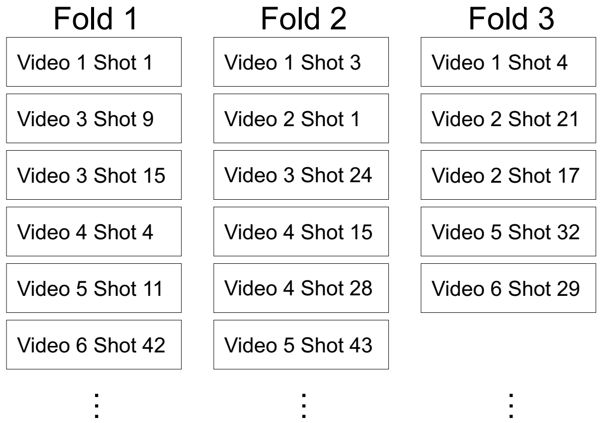
This paper shows that commonly used classifier performance evaluation techniques need special care when applied to multimedia classification. Much multimedia data is sequential in nature. For example, popular music has a verse and a chorus, in a multimedia presentation the slides are designed with a story in mind and in video data there is the temporal ordering of shots. This paper will show that sequential data requires extra effort to accurately estimate classification performance. As an instantiation of sequential multimedia data, we will focus on classifier estimation of semantic concept detectors in video. However, the described techniques readily apply to other types of data that share a sequential ordering.

The organization of this paper is as follows. The next section revisits standard classifier evaluation techniques. Then, section 3 introduces our cross-validation method for video classification. In section 4 the experimental setup is discussed followed by the results in section 5, and the conclusions in section 6.

2. CLASSIFIER PERFORMANCE EVALUATION

The error estimation of a classifier not only provides a qualitative assessment of the classifier, it also influences classifier performance. The estimation of the performance of the classifier over different features determines which features are used, and which features might be left out. Furthermore, several classifiers require parameters, which are tuned by maximizing the estimated performance over various settings. For example in a video classification task, Snoek et al. [10] use the estimated classifier performance to select the best low level features. Furthermore, they find the best parameters for a Support Vector Machine (SVM) [13] by minimizing the estimated classifier error. In their framework, inaccurate classifier performance estimation might result in choosing the wrong features, or in sub-optimal parameter settings. Hence, classifier error estimation affects the qual-

MM'06, October 23–27, 2006, Santa Barbara, California, USA.
ACM 1-59593-447-2/06/0010.



Fold 1	Fold 2	Fold 3
Video 1 Shot 1	Video 1 Shot 3	Video 1 Shot 4
Video 3 Shot 9	Video 2 Shot 1	Video 2 Shot 21
Video 3 Shot 15	Video 3 Shot 24	Video 2 Shot 17
Video 4 Shot 4	Video 4 Shot 15	Video 5 Shot 32
Video 5 Shot 11	Video 4 Shot 28	Video 6 Shot 29
Video 6 Shot 42	Video 5 Shot 43	
⋮	⋮	⋮

Figure 1: An example of partitioning a video set by using shot based 3-fold cross-validation.

ity of the classifier.

Estimating the classification error is done by training the classifier on one set, and testing the classifier on a different set. Thus, a straightforward approach to classifier performance estimation is by keeping part of the available data in an unseen hold-out set. This hold-out set should be as large as possible, to accurately represent the class variation that may be expected. However, keeping a large part of the data from the training set hinders the classification performance. The advantage of using a hold-out set is that the test-data is completely separate from the training data. However, keeping a large training set to obtain a good classifier, counteracts a large hold-out set for accurate error estimation.

In contrast to having a single hold-out set, the cross-validation method rotates the hold-out set over all available data. Cross-validation randomly splits the available data in X folds, where each of these X folds is once used as a hold-out set. The error estimations of all folds are averaged, yielding the classifier error. The cross-validation procedure may be repeated R times, to minimize the effect of the random partitioning. An example of cross-validation for a video set is shown in figure 1. The advantage of using cross-validation is the combination of a large training set with several independent test sets. Therefore, cross-validation is the standard procedure for classification error estimation [3, 4].

3. CROSS-VALIDATION IN VIDEO CLASSIFICATION

Machine learning is heavily used in semantic video indexing. The aim of semantic video indexing is detecting all relevant shots in a dataset to a given semantic category. Some examples of semantic concepts are *Airplane, Car, Computer Screen, Hu Jintao, Military Vehicle, Sports*. Indexed semantic concepts provide a user with tools to browse, explore, and find relevant shots in a large collection of video. With growing digital video collections, there is a need for automatic concept detection systems, providing instant access to digital collections. Therefore, machine learning techniques are vital to automatic video indexing.

In a video classification task, a shot is often the granularity of interest [8, 10]. However, a video document is the end result of an authoring process [10], where shots are used to convey a message. For example, a topic in news video, may consist of several shots, as shown in figure 2. Hence, a semantic concept might span several shots, while a video classification task is oriented towards single shots.

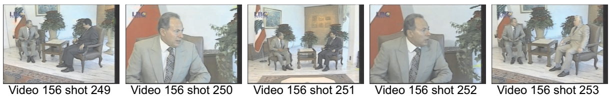
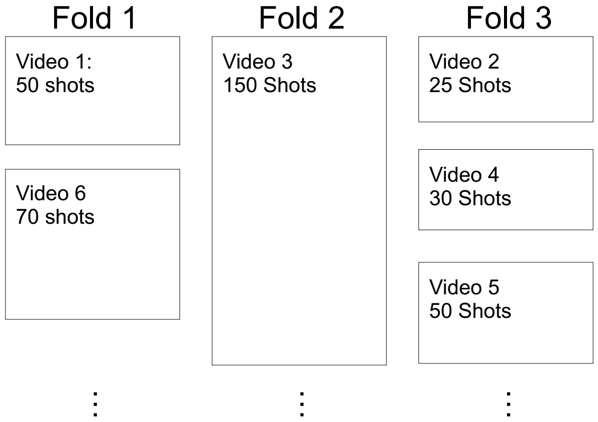


Figure 2: An example of narrative structure in video: five consecutive shots showing an interview with Lebanese President Émile Lahoud.

The mismatch between the granularity of the classification task and the granularity of the semantic concept requires special care in estimating classifier performance. Consider figure 2, and note the high similarity between shot 250 and shot 252. The similarity between these two shots can be expected, since they are part of the same narrative structure. However, the classification task focuses on single shots, and does not take this semantic relation between shots into account. Therefore, the common practice [8, 10] of estimating classifier performance by cross-validation on shots is biased. Cross-validation on shots will mix shots in a single topic to different folds while randomly partitioning the data. Thus, nearly identical shots will leak through to the rotating hold-out set. This leaking of near identical information creates a dependency between the training set and the hold-out set, which will manifest in too optimistic estimates for classifier performance. Moreover, if cross-validation is used for parameter tuning, the parameters will be biased towards near-duplicate data and might consequently fail to find the best classifier parameters for true independent hold-out data. Therefore, the sequential nature of video data should be taken into account when estimating classifier performance.

In order to preserve the semantic relation between shots in a topic, we propose an episode-constrained version of cross-validation. In contrast to a shot based partitioning of the video data, an episode-constrained partitioning treats a whole video episode as an atomic element. With videos as atomic elements, all shots in a video are kept together, preventing the leaking of near-identical shots to the hold-out set. Where the traditional method randomly distributes



Fold 1	Fold 2	Fold 3
Video 1: 50 shots	Video 3 150 Shots	Video 2 25 Shots
Video 6 70 shots		Video 4 30 Shots
		Video 5 50 Shots
⋮	⋮	⋮

Figure 3: An example of a partitioning a video set by using episode-constrained 3-fold cross-validation.

shots, our method randomly distributes videos. An example of episode-constrained cross-validation for a video set is shown in figure 3. Besides preventing the mixing of identical shots with the hold-out set, the episode-constrained method retains relations between shots. For example, in commercials a relation between shots is present, where both the content of the shots and the co-occurrence of the shots are identical. Therefore, the episode-constrained version of cross-validation creates truly independent hold-out data, and will yield more accurate performance estimates of video concept classification.

4. EXPERIMENTAL SETUP

In order to compare the episode-constrained version of cross-validation with the shot based version of cross validation, both methods are evaluated on a large corpus of news video. The evaluation was performed using the challenge problem for video indexing [11]. The challenge problem provides a benchmark framework for video indexing. The framework consist of visual features, text features, classifier models, a ground truth, and classification results for 101 semantic concepts[1] on 85 hours of international broadcast news data, from the TRECVID 2005/2006 benchmark [8]. The advantage of using the challenge framework is that the framework provides a standard set of features to the TRECVID data. Furthermore, the framework is well suited for our experiment, since there are a large number or shots, i.e. close to 45.000, and an abundance of semantic concepts.

The challenge data comes with a training set consisting of the first 70% of the video data, and a hold-out set containing the last 30% of the data. We use the training set for training a k-nearest neighbor classifier (kNN) [3]. The features we use are the visual features [12] that are provided with the framework. As the classifier performance measure we adopt average precision from the challenge framework. Average precision is a single-valued measure that summarizes the recall-precision curve.

5. RESULTS

The focus of the experiment is on comparing episode-constrained cross-validation versus shot based cross-validation. To this end, we use both cross-validation methods to estimate the best value for k for a kNN classifier, where $k \in \{1, 2, 3\}$. To evaluate the results, we computed the same parameters for k on the hold-out set. The results are displayed in figure 4, where the best result for each method and the hold-out set is emphasized.

Note that we are evaluating the effect of using a different strategy for creating random permutations of the training data. Since there is no need to create permutations of the hold-out set, there is only one column required for hold-out data in figure 4.

The first thing that is striking about the results in figure 4, is the discrepancy between methods in selecting the best classifier parameter. The shot based cross-validation method selects $k = 3$ for 97 out of 100 concepts. Moreover, the estimated average precision scores for most concepts are disproportional high compared to the scores that are obtained on the hold-out set. Consider for example the concept *Aircraft*. The shot based cross-validation predicts a score of 0.393 where the best parameter is $k = 3$ neighbors. In contrast, the episode-constrained cross-validation

	Shot Based			Episode Constrained			HoldOut		
	NN1	NN2	NN3	NN1	NN2	NN3	NN1	NN2	NN3
Aircraft	0.225	0.221	**0.393**	0.195	**0.203**	0.091	0.131	**0.132**	0.096
I. Allawi	0.271	0.248	**0.420**	**0.164**	0.153	0.132	**0.003**	0.001	0.000
Anchor	0.861	0.856	**0.910**	0.528	**0.538**	0.469	0.645	**0.650**	0.552
Animal	0.456	0.465	**0.566**	**0.393**	0.391	0.240	**0.481**	0.478	0.377
Y. Arafat	0.180	0.166	**0.326**	**0.174**	0.165	0.114	**0.192**	0.185	0.085
Basketball	0.384	0.355	**0.548**	0.053	0.055	**0.112**	**0.038**	0.036	0.036
Beach	0.233	0.211	**0.416**	0.126	0.127	**0.282**	0.044	0.036	**0.129**
Bicycle	0.436	**0.454**	0.427	**0.757**	0.750	0.588	0.504	**0.508**	0.056
Bird	**0.783**	0.779	0.741	**0.755**	0.753	0.638	0.887	**0.893**	0.847
Boat	0.289	0.280	**0.406**	0.276	**0.285**	0.131	0.225	**0.231**	0.094
Building	0.308	0.304	**0.458**	**0.269**	0.262	0.207	0.243	**0.251**	0.208
Bus	0.021	0.021	**0.186**	**0.030**	0.030	0.013	**0.047**	0.046	0.015
G. Bush jr.	0.310	0.290	**0.449**	**0.177**	0.172	0.121	0.034	**0.035**	0.019
G. Bush sr.	0.134	0.116	**0.344**	0.022	0.021	**0.171**	**0.000**	0.000	0.000
Candle	0.012	0.019	**0.295**	0.001	0.002	**0.011**	**0.030**	0.029	0.027
Car	0.378	0.372	**0.489**	**0.370**	0.359	0.205	**0.283**	0.282	0.161
Cartoon	0.748	0.692	**0.880**	0.780	**0.785**	0.781	**0.267**	0.258	0.232
Chair	0.274	0.275	**0.550**	0.252	**0.279**	0.241	**0.319**	0.317	0.286
Charts	0.491	0.476	**0.689**	0.367	0.339	**0.415**	0.393	**0.398**	0.348
B. Clinton	0.053	0.054	**0.335**	0.001	0.001	**0.007**	0.139	**0.148**	0.104
Cloud	0.158	0.150	**0.332**	0.137	**0.146**	0.067	0.130	**0.135**	0.088
Corporate leader	0.167	0.173	**0.355**	0.054	**0.056**	0.041	0.017	**0.017**	0.011
Court	0.122	0.125	**0.399**	0.060	0.063	**0.139**	0.208	**0.210**	0.105
Crowd	0.401	0.388	**0.553**	**0.373**	0.362	0.294	**0.391**	0.385	0.375
Cycling	0.459	**0.471**	0.456	0.782	**0.788**	0.521	0.629	**0.635**	0.074
Desert	0.131	0.131	**0.350**	0.074	**0.085**	0.068	0.072	**0.081**	0.047
Dog	0.383	0.386	**0.598**	0.400	0.401	**0.403**	0.298	**0.303**	0.174
Drawing	0.733	0.704	**0.786**	0.389	**0.389**	0.288	**0.210**	0.188	0.171
Drawing & Cartoon	0.746	0.737	**0.750**	**0.499**	0.451	0.324	**0.254**	0.246	0.210
Duo-anchor	0.523	0.525	**0.688**	0.092	**0.104**	0.072	**0.355**	0.313	0.353
Entertainment	0.604	0.588	**0.694**	**0.501**	0.471	0.414	**0.361**	0.345	0.265
Explosion	0.162	0.143	**0.414**	0.059	**0.060**	0.032	**0.087**	0.083	0.054
Face	0.914	0.911	**0.953**	**0.896**	0.895	0.839	0.865	**0.867**	0.810
Female	0.363	0.356	**0.575**	**0.129**	0.126	0.120	0.083	**0.086**	0.051
Fire weapon	0.182	0.177	**0.388**	0.092	0.108	**0.111**	0.041	0.041	**0.045**
Fish	0.545	0.535	**0.655**	0.437	0.434	**0.452**	0.848	**0.861**	0.597
Flag	0.278	0.255	**0.475**	0.093	0.091	**0.098**	0.046	**0.050**	0.025
Flag USA	0.355	0.334	**0.516**	**0.202**	0.187	0.122	0.048	**0.050**	0.021
Food	0.487	0.466	**0.601**	**0.375**	0.362	0.180	**0.374**	0.360	0.328
Football	0.201	0.206	**0.351**	**0.170**	0.165	0.168	**0.100**	0.097	0.017
Golf	0.531	0.547	**0.554**	**0.254**	0.246	0.231	**0.082**	0.070	0.076
Government building	0.174	0.164	**0.374**	0.053	0.056	**0.090**	0.015	**0.027**	0.026
Government leader	0.375	0.367	**0.529**	**0.249**	0.249	0.212	0.155	**0.159**	0.108
Graphics	0.488	0.484	**0.691**	0.361	**0.374**	0.319	**0.481**	0.480	0.437
Grass	0.417	0.410	**0.558**	0.283	**0.290**	0.193	**0.127**	0.119	0.055
H. Nasrallah	0.801	0.800	**0.817**	**1.000**	**1.000**	**1.000**	0.007	**0.008**	0.003
Horse	0.506	0.496	**0.594**	0.438	**0.442**	0.225	0.001	**0.001**	0.000
Horse racing	0.367	0.349	**0.566**	**0.103**	0.103	0.085	0.001	**0.001**	0.000
House	0.112	0.115	**0.206**	**0.037**	0.034	0.010	0.057	**0.063**	0.006
H. Jintao	0.334	0.334	**0.491**	0.087	0.086	**0.203**	**0.030**	0.021	0.007
Indoor	0.737	0.725	**0.820**	**0.537**	0.537	0.486	**0.573**	0.572	0.558
J. Kerry	0.184	0.186	**0.267**	0.041	**0.041**	0.027	0.000	0.000	**0.006**
E. Lahoud	0.676	0.671	**0.824**	0.681	0.702	**0.709**	**0.304**	0.289	0.220
Male	0.369	0.372	**0.591**	0.146	0.147	**0.152**	0.075	**0.075**	0.039
Maps	0.552	0.555	**0.713**	0.437	**0.450**	0.303	0.461	**0.468**	0.367
Meeting	0.335	0.323	**0.478**	0.272	**0.273**	0.205	0.143	**0.150**	0.145
Military	0.307	0.295	**0.458**	0.209	**0.216**	0.184	0.159	0.160	**0.160**
Monologue	0.348	0.329	**0.457**	**0.216**	0.202	0.093	**0.082**	0.079	0.047
Motorbike	0.801	0.800	**0.896**	0.668	0.668	**0.671**	0.008	**0.008**	0.002
Mountain	0.336	0.340	**0.475**	**0.305**	0.301	0.203	**0.215**	0.210	0.187
Natural disaster	0.140	0.140	**0.277**	**0.148**	0.144	0.062	0.107	**0.113**	0.047
News paper	0.649	0.638	**0.714**	**0.384**	0.378	0.355	0.301	0.304	**0.330**
Night fire	0.064	0.093	**0.359**	0.009	0.010	**0.023**	**0.412**	0.403	0.046
Office	0.162	0.165	**0.341**	**0.075**	0.071	0.065	**0.091**	0.088	0.041
Outdoor	0.730	0.719	**0.799**	**0.690**	0.676	0.610	**0.678**	0.676	0.666
Overlayed text	0.751	0.737	**0.814**	**0.657**	0.646	0.593	**0.640**	0.633	0.568
People	0.947	0.948	**0.981**	0.931	**0.933**	0.903	0.914	**0.916**	0.882
People marching	0.164	0.141	**0.316**	0.117	**0.120**	0.044	0.137	**0.140**	0.128
Police/security	0.070	0.072	**0.212**	0.090	**0.093**	0.027	**0.136**	0.117	0.084
C. Powell	0.214	0.246	**0.389**	0.004	0.001	**0.016**	**0.009**	0.008	0.003
Prisoner	0.053	0.052	**0.282**	0.006	0.005	**0.011**	**0.188**	0.120	0.149
Racing	0.009	0.006	**0.258**	0.001	0.001	**0.013**	0.003	**0.003**	0.001
Religious leader	0.166	0.167	**0.240**	0.115	0.111	**0.122**	**0.008**	0.008	0.003
River	0.482	0.501	**0.708**	0.500	**0.500**	0.354	**0.035**	0.023	0.009
Road	0.368	0.363	**0.493**	**0.316**	0.311	0.201	**0.201**	0.192	0.151
Screen	0.266	0.259	**0.383**	0.181	**0.189**	0.083	**0.148**	0.146	0.102
A. Sharon	0.251	0.251	**0.409**	0.389	0.389	**0.405**	0.003	0.003	**0.005**
Sky	0.466	0.459	**0.606**	**0.414**	0.411	0.341	0.340	**0.342**	0.329
Smoke	0.239	0.248	**0.417**	**0.162**	0.140	0.098	0.210	**0.228**	0.130
Snow	0.229	0.237	**0.435**	0.248	**0.256**	0.169	**0.217**	0.206	0.145
Soccer	0.581	0.568	**0.643**	**0.489**	0.488	0.346	**0.750**	0.750	0.459
Split screen	0.731	0.721	**0.799**	0.338	**0.342**	0.340	**0.388**	0.374	0.352
Sports	0.503	0.496	**0.586**	0.349	**0.349**	0.265	**0.242**	0.232	0.160
Studio	0.833	0.816	**0.888**	**0.572**	0.566	0.511	**0.643**	0.643	0.617
Swimming pool	0.441	0.346	**0.635**	**0.187**	0.186	0.187	0.082	**0.082**	0.023
Table	0.284	0.248	**0.400**	0.225	**0.241**	0.167	0.029	0.024	**0.031**
Tank	0.068	0.061	**0.305**	**0.187**	0.168	0.105	**0.008**	0.004	0.004
Tennis	0.678	0.681	**0.777**	0.647	0.634	**0.649**	**0.439**	0.408	0.401
T. Blair	0.336	0.336	**0.616**	0.403	**0.404**	0.288	0.006	0.004	**0.061**
Tower	0.246	0.242	**0.396**	**0.260**	0.250	0.129	0.132	**0.133**	0.039
Tree	0.316	0.317	**0.477**	0.284	**0.293**	0.232	0.105	**0.110**	0.053
Truck	0.102	0.104	**0.268**	0.069	**0.072**	0.020	**0.047**	0.045	0.027
Urban	0.409	0.390	**0.516**	**0.354**	0.342	0.224	**0.246**	0.238	0.220
Vegetation	0.303	0.300	**0.471**	**0.292**	0.287	0.198	**0.159**	0.158	0.095
Vehicle	0.397	0.388	**0.508**	**0.368**	0.359	0.213	0.312	**0.313**	0.208
Violence	0.460	0.455	**0.559**	**0.441**	0.439	0.317	0.282	**0.284**	0.260
People walking	0.431	0.418	**0.545**	**0.348**	0.336	0.285	**0.313**	0.313	0.287
Waterscape	0.437	0.419	**0.555**	**0.473**	0.466	0.270	0.361	**0.370**	0.210
Waterfall	0.320	0.281	**0.439**	**0.112**	0.111	0.083	0.502	**0.502**	0.231
Weather	0.246	0.228	**0.553**	0.228	0.222	**0.273**	0.000	**0.311**	0.219
Mean	**0.380**	**0.372**	**0.524**	**0.300**	**0.298**	**0.251**	**0.228**	**0.229**	**0.175**

Figure 4: Classification results in average precision for 100 concepts. Results for the k-nearest neighbor classifier, $k \in \{1, 2, 3\}$, are given for both partitioning strategies, and for hold-out data. The emphasized numbers represent the best score for each set. Note that episode-constrained cross-validation provides a more accurate estimation of classifier performance.

[1] We did not evaluate the concept *baseball*, since all the examples in the training set of this concept are found in a single video.

	Shot Based	Episode-Constrained
Training set	0.525	0.309
Hold-out set	0.188	0.221

Table 1: The mean performance over all concepts, using the estimated parameters as selected by each method.

predicts a score of 0.203 where the best parameter is $k = 2$ neighbors. Verifying the classification performance of *Aircraft* on hold-out data shows an average precision of 0.096 for $k = 3$ and a score of 0.132 for $k = 2$. Note that the performance estimate of episode-constrained cross-validation is not only closer to the hold-out performance, it also selects the best classifier parameter.

In table 1, we present the mean performance over all concepts, for both cross-validation methods. We show the estimated results on training data, and the results on hold-out data. The classifier parameter, the number of k-nearest neighbors, is tuned by selecting the maximum performance according to the cross-validation method at hand.

In analyzing table 1, we focus on two points: 1) the accuracy in estimating classifier performance and 2) the final classification performance. Starting with point 1, we consider the difference between the estimated performance on training data, and the reported performance on hold-out data. For shot based cross-validation there is considerable variation between the estimated performance on training data and the performance on hold-out data. Specifically, the performance estimate is 0.337 too optimistic. In contrast, for episode-constrained cross-validation the difference between training data and hold-out data is only 0.088. This clearly shows that the estimated performance of the episode-constrained cross-validation is more accurate than the performance estimate based on shots. Continuing with point 2, we compare the performance on hold-out data for both methods. It turns out that the episode-constrained method outperforms the shot based method by 17%. Analyzing the hold-out results per-concept, shows that episode-constrained cross-validation yields equal or better results for 93 out of 100 concepts, and it gives better results for 67 concepts. The shot based method gives the best results for 7 concepts. These results on hold-out data show that parameter tuning is considerably more accurate when using episode-constrained cross-validation.

6. CONCLUSIONS

In this paper, we compare two methods of cross-validation for estimating classification performance for semantic concept detection in video. The traditional method of cross-validation is based on shots, where we propose a method based on whole videos. This episode-constrained method for cross-validation prevents the leaking of similar shots to the rotating hold-out set. Experimental results show that the episode-constrained method yields a more accurate estimate of the classifier performance than the shot based method. Moreover, when cross-validation is used for parameter optimalization, the episode-constrained method is able to better estimate the optimal classifier parameters, resulting in higher performance on validation data compared to shot based (traditional) cross-validation.

7. ACKNOWLEDGMENTS

This research is sponsored by the BSIK MultimediaN project.

8. REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine*, 18(1):32–80, 2001.

[2] L.-Y. Duan, M. Xu, X.-D. Yu, and Q. Tian. A unified framework for semantic shot classification in sports video. *Transactions on Multimedia*, 7(6):1066–1083, 2005.

[3] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

[4] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *TPAMI*, 22(1):4–37, 2000.

[5] L. Lu, H. Jiang, and H. Zhang. A robust audio classification and segmentation method. In *ACM Multimedia*, pages 203–211, New York, NY, USA, 2001. ACM Press.

[6] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, and M. Barnard. Automatic analysis of multimodal group actions in meetings. *TPAMI*, 27(3):305–317, 2005.

[7] M. R. Naphade and T. S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *Transactions on Multimedia*, 3(1):141–151, 2001.

[8] NIST. TRECVID Video Retrieval Evaluation, 2001–2005. http://www-nlpir.nist.gov/projects/trecvid/.

[9] Y. Qi, A. Hauptmann, and T. Liu. Supervised classification for video shot segmentation. In *ICME*, July 2003.

[10] C. G. M. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders. The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *TPAMI*, 2006, in press.

[11] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM Multimedia*, 2006.

[12] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders. Robust scene categorization by learning image statistics in context. In *CVPR-SLAM*, 2006.

[13] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.