Integrated Classification of Audio, Video and Speech using Partitions of Low-Level Features

Edda Leopold[1], Jörg Kindermann[1], Gerhard Paaß[1],
Stephan Volmer[2], René Cavet[2],
Martha Larson[3], Stefan Eickeler[3], and Thorsten Kastner[4]

[1] Fraunhofer Institute for Autonomous intelligent Systems,
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
{Edda.Leopold, Joerg.Kindermann, Gerhard.Paass}@ais.fhg.de
[2] Fraunhofer Institute for Computer Graphics,
Fraunhoferstaße 5, 64283 Darmstadt, Germany
{volmer, rcavet}@igd.fhg.de
[3] Fraunhofer Institute for Media Communication
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
{larson, eickeler}@imk.fhg.de
[4] Fraunhofer Institute for Integrated Circuits
Am Weichselgarten 3, 91058 Erlangen, Germany
ksr@iis.fhg.de

Abstract. Multimodal documents are classified according to the IPTC annotation scheme. To this end usual text classification techniques are adapted to speech, video, and non-speech audio. To represent multimodal documents we apply the bag-of-words approach to speech-, audio-, and video-features. Word analogues are generated for the three modalities: sequences of phonemes or syllables for speech, 'video-words' based on low level color features, and 'audio-words' based on low-level spectral features for non-speech audio. Classification results based on video- or audio-words alone are comparable to those obtained on speech data.

1 Introduction

Content processing of speech, audio, and video data is one of the central issues of recent research in information management. During the last years new methods for the integrated classification of text, audio, video and voice information have been developed. The combination of features from different modalities should lead to an improvment of results. We use low-level features such as color histograms, spectral flatness, and phoneme sequences for the integrated classification of multimodal documents (A/V documents).

Support Vector Machines (SVM) have been applied successfully to text classification tasks [4, 2, 3, 7]. We adapt common SVM text classification techniques to A/V documents which contain speech, video, and non-speech audio data. To represent A/V documents we apply the bag-of-words approach (which is common to text classification). We generate word analogues for the three modalities:

sequences of phonemes or syllables for speech, "video-words" based on low level color features for video, and "audio-words" based on low-level spectral features for general audio.

2 The Motivation of our Approach

2.1 Quantitative Motivation

There is a trade-off between the semantic specificity of signs and their probability of occurrence. Signs with a very specific meaning usually are very rare. Therefore an in-depth analysis of the objects displayed in a picture and the relation between them may lead to sign aggregates which are semantically so specific that they occur very rarely. Such high-level features are likely not to occur in both test set and training set, which makes them useless for supervised classification algorithms. Using low-level features and a partition of the respective feature space for creating audio- and video-signs we are able to control the probability distributions of signs and adjust them properly to subsequent classification procedures.

Furthermore our approach of using low level features is in line with recent linguistic tendencies which prefer shallow parsing techniques rather than an in-depth semantic-syntactic analysis.

2.2 Philosophical Motivation

In his semiotic analysis of pictures Roland Barthes [1] distinguishes between the denoted message and a connoted message of a picture. In his view all imitative arts (drawings, paintings, cinema, theater) comprise two messages: a denoted message, which is the analogon itself, and a connoted message, which is the manner in which the society to a certain extent communicates what it thinks of it.

The denoted message of a picture is an analogical representation (a 'copy') of what is represented. For instance the denoted message of a picture which shows a person is the person itself. Therefore the denoted message of a picture is a simple agglutination of symbols which is not based on a true system of signs. It can be considered as a message without a code. The connotive code of a picture in contrast results from the historical or cultural experience of a communicating society. The code of the connoted system is constituted by a universal symbolic order and by a stock of stereotypes (schemes, colors, graphisms, gestures. expressions, arrangements of elements). [1]

The rationale behind the use of low-level video features is not to discover the denotated message of a video-artefact (whether it shows for instance a person or a car) but to reveal the implicit code which underlies its connoted message. The elements of the connoted code (i.e. schemes, colors, textures etc.) correspond to what is usually addressed as low-level features in the realm of image processing. Thus as proposed by [6] each element of a partition of feature space can be

considered as an (artificial) video-sign of the connoted code whereas the partition itself is the respective vocabulary of video-signs. We extend this idea to non-speech audio features, because determining the denotation of a non-speech audio artefact is usually impossible - not only for technical reasons but simply because it does not exist [9].

3 Feature Extraction

3.1 Speech

The acoustic signal was not separated into speech and non-speech segments. A continuous speech recognition system (CSR) was built using the ISIP (Institute for Signal and Information Processing) public domain speech recognition tool kit from Mississippi State University. It is a typical Hidden-Markov-Model-based system in which the basic acoustic models are phoneme models and consist of three states connected by forward transitions and self-transitions. At each state the probability that a state emitted the given feature vector is modeled by a probability density function composed of a mixture of Gaussians.

Our language model was syllable-based, built by stringing phoneme models together according to a pronunciation lexicon. The advantage of using a syllable-based language model instead of a word-based model is that it leads to reduction of vocabulary-size and results in less out-of-vocabulary errors which makes a domain dependent lexicon unnecessary. This is especially useful when the CSR is applied to a language which is highly productive at the morpho-syntactic level, like German in our case [5]. N-grams were constructed from the recognized syllables and phonemes in order to reach a level of semantic specificity which is comparable to that of words. The use of n-grams also makes it possible to adjust the linguistic units appropriate to the trade-off between semantic specificity and low probability of occurrence, which is especially important when document classes are small. [10]

3.2 Creating Visual and Acoustic Vocabularies

Most results presented in the paper were obtained by using a vocabulary which is drawn from 11 hours of video in october 2002. Vocabularies from the corpus itself and from January 2003 were used only to get an insight in the temporal variation of the visual code. Interestingly comparison of results based on the different vocabularies did not differ very much.

Video The generation of a visual vocabulary is done on a corpus of video data from recorded TV news broadcasts. In our tests, the video set for training had an approximate length of eleven hours. First the video data was split into individual frames. To reduce the huge amount of frames, only one frame per second of video material was selected . This could be done because of the high similarity between the neighboring frames. This reduces the frames count from approx. 500000 to

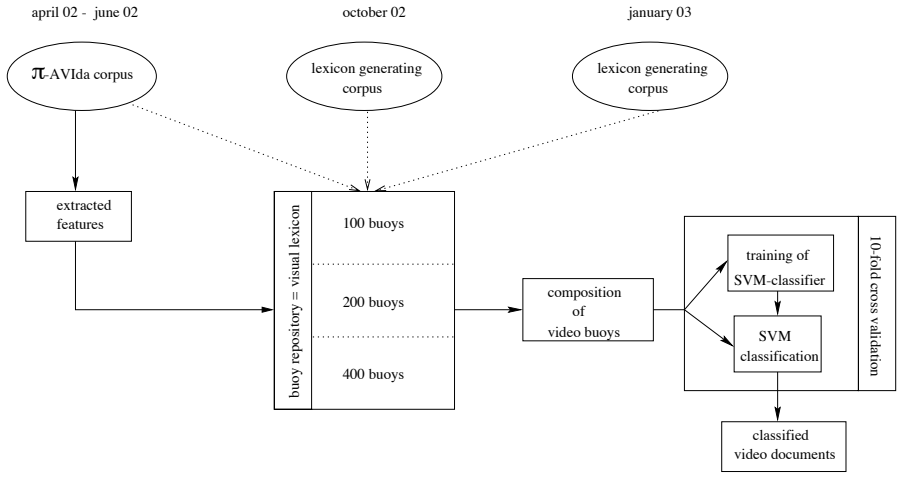

Fig. 1. Outline of our experimental design (video only) Most results presented in the paper were obtained by using a vocabulary which was drawn from the october 2002 data. Vocabularies from the corpus itself and from January 2003 were used only for comparison.

17342 frames. After that, the content feature descriptors were extracted on the reduced set of frames and the buoy generation method [11] is applied to each descriptor set.

Three low-level features were extracted from each key-frame: a histogram of 29 colors, a correlogram calculated on the basis of 9 major colors, and first and second moment of the distributions of each the 9 major colors. For each of these features vocabularies of different sizes (100, 200, 400, and 800) of video-buoys were generated. They represent visual vocabularies (or visual lexicons) of different sizes. A larger lexicon size implies more specific video-signs, each covering a smaller semiotic extension on average. Buoy sets were generated also from video data recorded at different time periods: April 2002 to July 2002 (this data set comprises the multimodal corpus described below i.e. the A/V scenes to be classified whereas the other sets were solely used for lexicon generation), October 2002, and January 2003. This was done in order to get an insight into the temporal variation of the visual lexicon.

Audio The low-level audio features that we considered were Audio-Spectrum-Flatness and Audio-Spectrum-Envelope as described in MPEG-7-Audio. Audio-Spectrum-Flatness was measured for 16 frequency bands ranging from 250 Hz to 4 kHz for every audio frame of 30 msec. The Audio-Spectrum-Envelope was calculated for 16 frequency bands ranging from 250 Hz to 4 kHz plus additional bands for the low-frequency (below 250 Hz) and high-frequency (above 4 kHz) signals. For both features a vocabulary of 1000 video-buoys was generated. We

are in the process of experimenting with different sizes and creation dates of the acoustic vocabulary. These experiments, however, have not yet been completed.

3.3 Mapping A/V Scenes to Audio- and Video-Words

The A/V stream is segmented manually into semantically coherent scenes that belong to one news story. The first step for representing the video scenes is the segmentation of the video stream into coherent units (shots). For each shot a representative picture is selected, the "key frame". The segmentation is done by algorithms monitoring the change of image over time. Two adjacent frames are compared and their difference is calculated. The differences are summed, and when the sum exceeds a given threshold a shot-boundary is detected, and the key frame of the shot is calculated. For each shot the three low-level features extracted from its key frame. Each of the three visual features is mapped to the nearest video-buoy in the respective visual vocabulary. The three buoy IDs are concatenated to form a *video-word*, which represents the shot.

Both acoustic feature vectors are mapped to the nearest audio-buoy of its respective vocabulary. This way every audio-frame of 30ms is represented by two audio-buoys, which are combined to form an *audio-word*.

Repeated sequences of *identical* video or audio-words are reduced to a single video- or audio-word. Resulting video- or audio-words are combined to form sequences (n-grams) up to a length of $n \leq 5$. Further processing follows the usual bag-of-words representation which is commonly applied for text classification: For each scene (and for each of the three modalities: video, audio, and speech) a type-frequency vector is generated which contains the number of occurrences for each n-gram. These type-frequency vectors are concatenated and span a product space for the integrated multi-modal classification.

The major difference between audio and video-processing is that audio-words (for the time being) were generated for each audio-frame (30 msec) whereas video-words were created for every shot, which may last up to 3 sec. The poor results for audio classification reported below show that this technique has to be improved and that audio-words have to be created for temporally larger units. Respective experiments are in progress but have not been finished yet.

4 The Multimodal Corpus

The corpus consists of 693 multimodal (A/V)-documents. Segmentation into semantically coherent scenes (documents) and semantic annotation according to the categorization scheme of the International Press Telecommunications Council (IPTC) (see http://www.iptc.org) was done manually. The data were obtained from two different German news broadcast stations: N24 and n-tv. Document length ranges between 30 sec. and 3 minutes. The material from N24 consists of 353 A/V-documents and covers the period between May 15 and June 13, 2002 (including reports from the World Cup soccer tournament) in Korea and Japan, which can be considered as a semantically unique event, that does not appear in

the "vocabulary" obtained from tv recordings of November). The data from n-tv comprises 340 documents and covers the last seven days of April 2002. Table 1 shows the distribution of topic classes in the corpus. For convenience we added two classes "advertisement" and "jingle" to the 17 top level classes of the IPTC-categorization. The number of documents in the classes total more than 693, because some documents were attributed to two or three classes because of the ambiguity of their content. For example A/V-documents on the Israel-Palestine conflict often were categorized as belonging to both "politics" and "conflicts".

Note that the size of the classes varies considerably; "politics" comprises 200 A/V-documents whereas "religion" contains just 4. We only used the seven categories with more than 45 documents (shown in the right column of table 1) for classification experiments. This means that all documents of the other (small) categories were always in the set of counter-examples.

Table 1. Size of IPTC-classes in terms of number of documents. Only those classes, which contain more than 45 documents (right columns) were considered

category	docs.	category	docs.
religion	4	labour	49
social issue	6	economy	68
weather	8	conflicts	85
education	10	sports	91
science	13	advertisement	119
leisure	15	justice	120
environmental issue	17	politics	200
health	19		
culture	22		
jingles	22		
disaster	38		
human interest	40		

5 The Classification Procedure

Each video scene d_i is represented by its type-frequency vector

$$\mathbf{f}_i = \big(r_1 \cdot f(w_1, d_i), \ldots,\ r_n \cdot f(w_n, d_i)\big) \tag{1}$$

where r_j is an importance weight as described below, w_j is the j-th n-gram (or the j-th type generated by the visual vocabulary), and $f(w_k, d_i)$ indicates how often w_j occurs in the video scene d_i. Type-frequency vectors are normalized to unit length with respect to L_1. In subsequent tables the use of type-frequencies is indicated by "rel". The vector of logarithmic type-frequencies of a video scene

d_i is defined as

$$\mathbf{l}_i = \Big(r_1 \log\big(1 + f(w_1, d_i)\big), \ldots, r_n \log\big(1 + f(w_n, d_i)\big)\Big) \tag{2}$$

Logarithmic frequencies are normalized to unit length with respect to L_2. Other combinations of norm and frequency transformation were omitted because they appeared to yield worse results. In tables below the use of logarithmic type-frequencies is indicated by "log".

As there is a large number of possible n-grams in the scenes, a statistical test is used to eliminate unimportant ones. First it is required that each type must occur at least twice in the corpus. In addition the hypothesis that there is a statistical relation between the document class under consideration and the occurrence of a type w is investigated by a χ^2-statistic. A type w is rejected when its χ^2 statistic is below a threshold θ. The values of θ used in the experiments are $\theta = 0.001$ and $\theta = 1$.

Importance weights like the well-known inverse document frequency are used often used in text classification in order to quantify how specific a given type is to the documents of a collection. Here another importance weight, namely redundancy, is used. In information theory the usual definition of redundancy is maximum entropy ($\log N$) minus actual entropy. So redundancy is calculated as follows: consider the empirical distribution of a term over the documents in the collection and define the importance weight of type w_k by

$$r_k \quad = \quad \log N + \sum_{i=1}^{N} \frac{f(w_k, d_i)}{f(w_k)} \log \frac{f(w_k, d_i)}{f(w_k)}, \tag{3}$$

where $f(w_k, d_i)$ is the frequency of occurrence of term w_k in document t_i and N is the number of documents in the collection. The advantage of redundancy over inverse document frequency is that it does not simply count the documents that a type occurs in, but takes into account the frequencies of occurrence in each of the documents. Since it was observed in previous work [7] that redundancy is more effective than inverse document frequency, two experimental setting are considered in this paper: term frequencies $f(w_k, d_i)$ are multiplied by r_k as defined in equation 3 (denoted by "+" at column "r" in subsequent tables); or term frequencies are left as they are: $r_k \equiv 1$ (denoted by "–").

It is well known that the choice of the kernel function is crucial to the efficiency of support vector machines. Therefore the data transformations described above were combined with the following different *homogeneous* kernel functions:

- Linear kernel (L)
  $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$
- 2nd and 3rd order polynomial kernel (P(d))
  $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^d \qquad d = 2, 3$
- Gaussian rbf-kernel (R(γ))
  $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma ||\mathbf{x}_i - \mathbf{x}_j||} \qquad \gamma = 0.2, 1, 5$
- Sigmoidal kernel (S)
  $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\mathbf{x}_i \cdot \mathbf{x}_j)$

In some of the experiments (see table 8) these *homogeneous* kernel functions were combined to form *composite* kernels, which use different kernel functions for each modality (for example L for speech, R(1) for video and P(3) for audio). Formally a composite kernel is defined as follows: Let the input space consist of L_s speech attributes, L_v video attributes, and L_a audio attributes, which are ordered in such a way, that dimension 1 to L_s correspond to speech attributes, dimensions $L_s + 1$ to $L_s + L_v$ correspond to video attributes, and dimensions $L_s + L_v + 1$ to $L_s + L_v + L_a$ correspond to audio attributes. Let $\pi_l^k(\cdot)$ be the projection from the input space to its subspace spanned by dimensions k to l. A composite kernel that uses kernel K_1 for speech, K_2 for video and K_3 for audio is defined as

$$\begin{aligned} K_{K_1,K_2,K_3}(\mathbf{x}_i, \mathbf{x}_j) = & K_1\Big(\pi^1_{L_s}(\mathbf{x}_i), \pi^1_{L_s}(\mathbf{x}_j)\Big) \\ & + K_2\Big(\pi^{L_s+1}_{L_s+L_v}(\mathbf{x}_i), \pi^{L_s+1}_{L_s+L_v}(\mathbf{x}_j)\Big) \\ & + K_3\Big(\pi^{L_s+L_v+1}_{L_s+L_v+L_a}(\mathbf{x}_i), \pi^{L_s+L_v+1}_{L_s+L_v+L_a}(\mathbf{x}_j)\Big) \end{aligned}$$

The idea behind the construction of composite kernels is that different semiotic and cognitive conditions for speech, video and audio imply different geometries in the respective factor spaces. I.e. we treat audio, video, and speech differently although we represent them in the same imput space. A kernel is called homogeneous kernel if $K_1 = K_2 = K_3$.

Table 2. Results of the classification on the basis of phoneme sequences.

category	n	r.	θ	tra.	kernel	F-score
justice	3	+	0.01	rel	R(5)	67.0
economy	2	-	1.00	log	R(1)	60.2
labour	2	+	0.01	log	S	83.6
politics	3	+	1.00	rel	R(5)	74.7
sports	2	+	1.00	rel	R(5)	84.8
conflicts	3	+	1.00	rel	R(5)	72.2
advertis.	2	-	1.00	log	P(3)	88.9

6 Results

The following tables show the classification results on the basis of the different modalities. A "+" in the column "r." indicates that the importance-weight redundancy is used, and "-" indicates that no importance weight is used. The values of the significance threshold θ used in our experiments are $\theta = 0.01$ and $\theta = 1$. The column "tra." indicates the frequency transformation that was used,

"log" stands for logarithmic frequencies with L_2-normalization and "rel" means relative frequencies (i. e. frequencies with L_1-normalization). The next column "kernel" indicates the kernel function: L is the linear kernel, S is a sigmoidal kernel, and $P(d)$ and $R(\gamma)$ denote the polynomial kernel and the rbf-kernel respectively. The last column shows the classification result in terms of the $F - score$, which is calculated as

$$F = \frac{2}{\frac{1}{prec} + \frac{1}{rec}},$$

where *rec* and *prec* are the usual definitions of the recall and precision [8]. For the classification a 1-to-n scheme was used, i.e. each class was classified against all other documents. All classification results presented in this section were obtained by tenfold crossvalidation, where the lexicon is held constant. This makes the results statistically reliable. Crossvalidation involving lexicon generation is unnecessary because the data set used for lexicon generation is different from the multimedia corpus.

6.1 Classification Based on Speech

The results on phoneme-based classification are shown in tables 2 and 3. The classifier uses sequences of up to 5 phonemes. This nearly reaches the average word-length in German which is about 5.3 phonemes.

If syllables are used instead of phonemes, unigrams yield the best performance for all classes. This result is in line with earlier experiments on the classification of German spoken documents [10].

Surprisingly the identification of the class "environmental issues" is based on syllable trigrams. These units are in average as long as 1.5 words, since the average word-length in German is 1.9. Phoneme-sequences failed for identifying "environmental issues", obviously this class cannot be identified on the basis of short linguistic units.

Table 3. Results of the classification on the basis of syllable sequences.

category	n	r.	θ	tra.	kernel	F-score
justice	1	+	0.01	rel	R(5)	65.0
economy	1	+	1.00	rel	P(2)	59.3
labour	1	+	1.00	log	S	85.3
politics	1	+	0.01	rel	R(1)	74.7
sports	1	+	0.01	rel	R(5)	80.3
conflicts	1	+	1.00	rel	R(5)	73.5
advertis.	1	-	1.00	log	P(2)	85.0

6.2 Classification Based on Video

Table 4 shows results based on a set of 100 video buoys, table 5 those on 400 video buoys. In the case of a visual lexicon of 100 video buoys the units used for the classification are n-grams with a size varying from $n = 1$ to $n = 3$ (note that such a unit may last up to 5 seconds). This means that these units are built from one to three video-shots. Those categories that are classified on the basis of shot-unigrams show relatively poor performance. We therefore suppose that we have detected regularities in the succession of video-units, which reveal a kind of temporal (as opposed to spatial) video-syntax. Linear kernels never perform best for any category. Rbf-kernels seem to be the most appropriate for classification on the basis of video-words when a small set of video buoys is considered.

With more buoys to choose from, the performance increases significantly with the exception of the category 'labour'. The n-gram degree decreases, and there is a large variety of kernel functions. We attribute this to the fact that the semantic specificity of n-grams increases with n. As units from a larger vocabulary are on average semantically more specific than units from a smaller one, the specificity of video-words obtained from the larger vocabulary is compensated by a decrease of n-gram degree.

Table 4. Classification results based on a visual vocabulary of 100 video buoys

category	n	**r.**	θ	**tra.**	**kernel**	F
justice	2	−	1.00	rel	R(1.0)	49.796
economy	3	+	1.00	rel	R(5)	24.490
labour	1	−	0.01	rel	S	33.333
politics	3	−	1.00	rel	S	43.928
sports	3	+	1.00	rel	R(0.2)	46.739
conflicts	1	−	0.01	rel	R(1.0)	35.000
ads	2	+	0.01	rel	R(5.0)	84.581

Table 5. Classification results based on a visual vocabulary of 400 video buoys

category	n	**r.**	θ	**tra.**	**kernel**	F
justice	1	−	1.00	log	S	51.6
economy	1	+	1.00	log	L	39.3
labour	1	−	0.01	log	S	16.1
politics	2	−	1.00	log	P(3)	57.4
sports	2	+	1.00	log	S	49.7
conflicts	2	−	0.01	log	R(0.2)	44.7
ads	2	+	0.01	rel	R(5.0)	90.0

The effect of date of lexicon creation One may argue that using video buoys which were generated after the acquisition of the test corpus may be flawed because typical pictures cannot be present in video material obtained at a later interval of time. However our principal assumption was that the video buoys reveal a kind of implicit code, which is known to the individuals of a given society. The assumption of such a code implies that it is shared by the members of the society and functions as a means to convey (non-linguistic) information. To fulfil this communicative function the code may not vary too quickly. As can be seen in figure 3, experimental results with visual lexicons created at different times (summer 2002 and January 2003) did not show a consistent change in performance and support the assumption of independence from the date of lexicon acquisition.

From figure 3 one can see that an effect of a steady evolution of the visual semiotic system is reflected in the results of the classification. Categories justice and sports and to a lesser extent politics show a sharp decline of performance when the lexicon was drawn from the October material instead of the corpus itself. This can attributed to the fact that there were salient news in these categories at the time when the corpus was sampled, namely the socker world championship (sports) and a massacre at a German high school (justice).

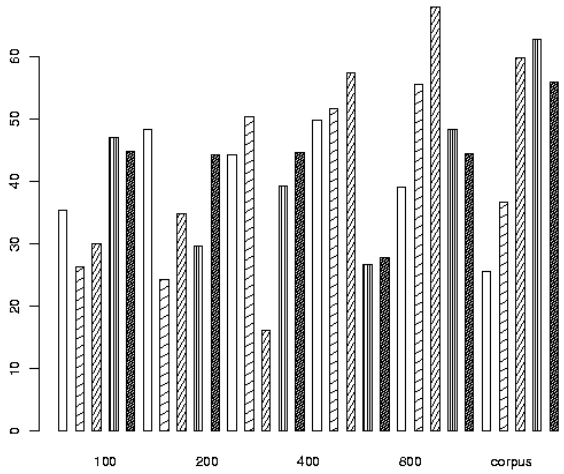

Fig. 2. Classification performance vs. vocabulary size. All vocabularies except corpus where obtained from October 2002. One can see that different classes show different behaviour when the size of vocabulary is changed. Note that the visual vocabulary which was obtained from the corpus itself (labeled as "corpus") does not yield outstanding classification results compared to the other

The results for the categories economy and conflicts are nearly independent from the creation date of the visual lexicon. These categories are communicated

by visual signs which seem to be temporarily invariant as far as they can be described by color based low-level features.

The category labour again plays a special role. On the video buoys obtained from corpus itself this category shows reasonable classification performance. However there is a poor temporal generalization. It might be that in the case of labour the artificially generated video-signs do not correspond to those visual signs which are actually used in the communicating society.

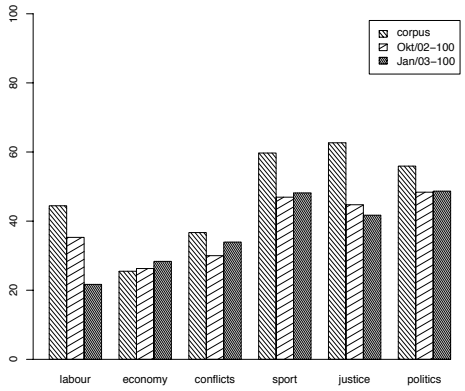

Fig. 3. Classification performance vs. date of vocabulary acquisition

6.3 Classification Based on General Audio

Classification on the basis of the audio-words is shown in table 6. The error rates are worse than those of the other two modalities (speech and video). Those categories which show a good performance mostly are classified on the basis of audio-word unigrams. This means that building sequences of audio-words in general does not improve the performance. Keeping in mind that audio-words are generated for every audio-frame of 30 msec., this suggests that there are only little regularities in the temporal combination of audio-units at this timescale. Performance may improve when audio-words are generated for larger time intervals. Polynomial-kernels seem to be most appropriate for classification on the basis of audio-words.

We have not yet experimented with different dates of lexicon acquisition and different sizes of acoustic vocabularies. This is ongoing work as well as the utilization of acoustic units spanning a larger interval of time.

Table 6. Results of the classification on the basis of sequences of audio-words

category	n	r.	θ	tra.	kernel	F-score
justice	1	-	0.01	rel	R(5)	37.57
economy	4	-	1.00	rel	P(3)	15.13
labour	1	-	1.00	rel	P(3)	13.33
politics	1	-	0.01	rel	R(5)	41.96
sports	1	-	1.00	rel	P(2)	28.57
conflicts	3	+	1.00	rel	P(3)	21.43
advertis.	1	-	0.01	log	S	90.60

6.4 Comparison of Results from Video, General Audio, and Speech

Figure 7 compares the error rates of all modalities directly. For most classes phoneme sequences yield best results. In some of the cases where phoneme sequences are comparatively bad, other modalities improve the picture.

Table 7. Comparison of the results on different modalities

category	phon.	syl.	video	audio	docs.
justice	**54.92**	53.10	49.80	37.57	120
economy	**50.00**	43.48	24.49	15.13	68
labour	**65.57**	61.82	33.33	13.33	49
politics	**64.91**	64.30	43.93	41.96	200
sports	50.20	**51.03**	46.74	28.57	91
conflicts	42.19	**43.36**	35.00	21.43	85
advertis.	90.46	88.98	84.58	**90.60**	119

Classification on the basis of audio-words yields worse results than the other two modalities (speech and video). There is however one big exception: advertisement. The generally superb rates of the category "advertisement" are likely to be caused by the following: The shot duration in commercial spots is generally very short, and therefore it exhibits a temporal visual syntax different from other categories. Furthermore the overall energy in the audio spectrum tends to be considerably higher than normal: advertisements sound more "intense" to the human ear than other broadcasts. Another aspect is that commercials are broadcast repeatedly. This means that in some cases identical spots tend to be present in both test and training data.

6.5 Integrated Classification using Composite Kernels

We explored the use of composite SVM kernels which apply different input space geometries for different modalities. Table 8 shows preliminary results.

Table 8. Results of the classification using composite kernels. The vocabularies were created from 100 video-buoys and 1000 audio-buoys. Speech was represented in terms of syllables

			n-gram			kernel			
cat.	r.	tra.	a	s	v	a	s	v	F
justice	+	rel	1	2	2	S	P(2)	P(2)	50.8
econ.	+	log	1	1	1	P(2)	S	R(1)	45.2
labour	-	log	1	1	1	P(3)	S	R(1)	58.2
politics	+	rel	1	1	1	P(3)	R(1)	R(1)	62.0
sport	+	rel	1	1	1	P(2)	R(1)	R(1)	49.3
confl.	+	rel	1	1	2	P(2)	R(1)	R(1)	45.7
adv.	-	rel	1	1	1	P(2)	R(1)	S	88.6

The columns labeled "a", "s", and "v" show the n-gram degrees and kernel parameters for audio, speech, and video sections of the sign frequency vectors. Comparison with table 7 reveals that results are as yet inferior to the best results based on single modalities. The results presented in table 8 have been obtained with a visual inventory of 100 video-buoys.

6.6 Integrated Classification using Homogeneous Kernels

Further experiments were carried out with homogeneous kernels and different sizes of the visual vocabulary combined with speech. It turned out that 400 buoys yield the best results (see table 9). Comparison with table 7 shows a 5% increase in F-score with exception of the class "labour". This class has already shown poor performance when classification was based on video alone (see table 4). In this case the video information seems to be misleading and reducing the good results on syllables and phonemes alone.

Table 9. Results of the classification with sequences of video-words (400 buoys) and syllables

category	n (video)	n (speech)	r.	θ	tra.	kernel	F-score
justice	1	1	+	1.00	log	P(2)	62.7
economy	2	2	-	1.00	log	S	59.2
labour	1	1	-	0.01	log	S	78.0
politics	2	1	+	1.00	log	S	65.3
sports	2	2	+	1.00	log	S	86.6
conflicts	1	2	+	1.00	log	S	71.0
advertis.	1	2	+	1.00	log	S	93.4

The results on video and speech indicate that the adjustment of lexicon sizes in the different modalities is crucial to a successful integrated classification of multimodal documents. We hope that ongoing experiments with different settings in the acoustical domain will lead to a further improvement of integrated classification.

7 Conclusion

Audio- and video-words constructed from low-level features provide a good basis for the classification of A/V-documents. The best F-scores on our difficult corpus range between 50% and 90%. Visual vocabularies generated as described in this paper are to a certain extent temporally stable. This allows to create a visual lexicon before the actual video classification is performed. The classification performance depends on the lexicon size. As units from a larger vocabulary are on average semantically more specific than units from a smaller one, the specificity of video-words obtained from the larger vocabulary is compensated by a decrease of n-gram degree. Using sequences of audio-words generated for every single audio-frame of 30 msec. yields poor classification performance (except for commercials). This suggests that there are no regularities in the temporal combination of audio units at this timescale. Performance may improve when audio-words are generated from larger time intervals. Composite kernels which induce different geometries on different modalities do not lead to an improvement of classification. However a proper adjustment of lexicon sizes of the different modalities is crucial to a successful integrated classification of multimodal documents. In this respect it is questionable if representation of the different modalities in a joint product space is the best solution to the problem of media integration.

8 Acknowledgment

This study is part of the project Pi-AVIda which is funded by the German ministry for research and technology (BMFT) (proj. nr. 107). We thank the Institute for Communication and Phonetics of the University of Bonn for contributing the BOSSII system and we thank Thorsten Joachims (Cornell University) who provided the SVM-implementation SVM^{light}.

References

1. Barthes, R.: Image, Music, Text. Noonday Press (1988)
2. Dumais, S., Platt, J., Heckerman, D., Sahami, M.: Inductive learning algorithms and representations for text categorization. In: 7th International Conference on Information and Knowledge Management (1998)
3. Drucker, H., Wu, D., Vapnik, V. Support vector machines for spam categorization. In: IEEE Transactions on Neural Networks **10** (1999) 1048–1054

4. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the Tenth European Conference on Machine Learning (ECML 1998) Springer Lecture Notes in Computer Science, Vol. 1398, Springer-Verlag (1998) 137–142
5. Larson, M., Eickeler, S., Paaß, G., Leopold, E., Kindermann, J.: Support Vector Machines for German Spoken Document Classification. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP) vol. 3, (2002) 1989–1992.
6. Leopold, E.: Artificial Semiotics. 10th International Congress of the German Society of Semiotics (DGS) (2002).
7. Leopold, E., Kindermann, J.: Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? Machine Learning, **46** (2002) 423–444
8. Manning, C., Schütze, D.: Foundations of Statistical Natural Language Processing. MIT Press (1999)
9. Nattiez, J.-J.: De la sémiologie à la musique. Université du Québec à Montreal, Montreal (1988)
10. Paaß, G., Leopold, E., Larson, M., Kindermann, J., Eickeler, S.: SVM Classification Using Sequences of Phonemes and Syllables. In: Elomaa T., Mannila H., Toivonen H. (eds.): Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2002), Springer (2002) 373–384
11. Volmer, S.: Fast Approximate Nearest-Neighbor Queries in Metric Feature Spaces by Buoy Indexing. In: Proc. 5th International Conference on Visual Information Systems, Hsinchu Taiwan (2002) 36–49