Exploiting document feature interactions for efficient information fusion in high dimensional spaces

Jana Kludas, Eric Bruno and Stephane Marchand-Maillet
Viper group – CS Department/CUI
University of Geneva, Switzerland
Email: [kludas,bruno,marchand]@cui.unige.ch

***Abstract*— Information fusion, especially for high dimensional multimedia data, is still an open research problem. In this article, we present a new approach to target this problem. Feature information interaction is an information-theoretic dependence measure that can determine synergy and redundancy between attributes, which then can be exploited with feature selection and construction towards more efficient information fusion. This also leads to improved performances for algorithms that rely on information fusion like multimedia document classification. We show that synergetic and redundant feature pairs require different fusion strategies for optimal exploitation. The approach is compared to classical feature selection strategies based on correlation and mutual information.**

***Keywords*— multimodal information fusion, feature selection, feature information interaction**

I. Introduction

Multimedia data processing received, in the last decade, a lot of attention by research communities due to the "multimediatisation" of the WWW as well as of private and professional data collections. Due to the multimodal nature of their input data, multimedia indexing, retrieval and classification systems have in their cores an essential need for information fusion. Over the last decades, the fusion of information with its various application areas has established itself as an independent research area but a general theoretic framework to describe information fusion systems is still missing [9]. The still limited understanding of how fusion works and by what it is influenced is probably a reason why multimedia retrieval is up to today staying behind expectations. For example, in the INEX 2006 Multimedia Task [15] text-only based runs outperformed multimedia-based ones. In INEX 2007, this did not change. The results of the TrecVid evaluation workshop, where the top ranked runs all use multimedia data are an exception because the noisy speech transcript weakens the textual modality.

The work so far done on information fusion in multimedia settings may be divided into two main directions: (1) fusion of independent or complementary information by assuming or creating independence and (2) fusion of dependent information by exploiting their inhered statistical dependencies. Both approaches have been applied in multimedia processing problems equally successful - neither of this approaches is in general superior.

To circumvent the "curse of dimensionality" that hinders information fusion in high dimensional spaces characteristic to multimedia data, preprocessing techniques like feature selection and construction are often incorporated. Those algorithms try to determine relevant features and modalities towards the fusion goal with "feature relevance" meaning "optimality" being still under discussion [8].

Aligned to the second idea, we present an information fusion approach that is based on the prediction of optimal feature subsets out of all modalities using as a measure feature information interactions. It is an entropy-based dependence measure that is superior to traditional dependence ones due to its consistent definition, its global application to the whole feature set (instead of only looking at pairwise dependencies) and its targeting at linear as well as higher order statistical dependencies. Under the light of this new definition of multivariate feature relations, we show how current machine learning algorithms including feature selectors do not treat the feature statistical dependencies right. They mostly apply a greedy or myopic strategy based on local, pairwise feature relations.

In section II, we present related work that is engaged in finding complex feature relationships for feature selection and information fusion. Next, the theory of feature information interaction is given in section III, including a discussion on how its different types, synergy and redundancy, can be exploited to improve fusion algorithms. Experiments on artificial data sets show what feature relationships can be detected with feature interactions and that they outperform pairwise dependence measures (section IV). Classification experiments with a multimedia data collection show the superiority of our approach compared to greedy feature selectors e.g. based on correlation and mutual information (section V).

II. Related Work

The importance of feature interactions for e.g. data mining, knowledge detection, dimensionality reduction and feature selection is an intense and fruitful field of research. For example, [4] gives a good introductory discussion about the relevance of attribute interactions for attribute construction, detection of Simpson's paradox, coping with attribute disjoints and why greedy attribute selection does not work well. As a solution, the researchers suggest genetic algorithms to conduct an efficient global search.

Another early paper ([10]) proposes to replace the greedy or so-called myopic feature selections in inductive learning with

the RELIEFF system. It learns top-down decision trees that capture conditional dependencies between attributes largely outperforms other machine learning approaches of that time like Naive Bayes, LFC (lookahead Feature Construction) and k-NN. A recent extension of this approach, the tree dependent component analysis (TCA) learns within-cluster dependencies and independencies between clusters [1].

In [8], wrapper methods for feature selection are presented that outperform filter approaches based on decision trees like RELIEFF. The main disadvantage of filter approaches is the ignorance of the effects of the feature selection on the induction algorithm performance. This is overcome by forward or backward searching in the feature subsets and evaluating their optimality by means of the induction algorithms accuracy.

More recent approaches include multidimensional projections to find complex logical feature relations (see e.g. [13]). Research on how to calculate information-theoretic quantities efficiently for large dimensions with having only a little number of training data can be found in [14].

Probably the most related work to ours is presented in [5]. In bio-informatics, they already use feature relations (synergistic gene pairs) to improve microarray-based classification, which needs simultaneous profiling of thousands of genes with various conditions. Since the problem definition is very similar to our multimedia document classification tasks (high dimensionality, small number of training data, low level features), we started to investigate its application in multimedia settings.

III. Feature Information Interaction

Interactions are a multivariate, information-theoretic based feature dependence measure [6], [7]. Before its introduction, there was no unifying definition of feature dependence in multivariate settings, but similar formulae have emerged independently in other fields from physics to psychology. Feature information interaction, or co-information as it was named in [2], is based on McGill's multivariate generalization of Shannon's mutual information. It describes the information that is shared exclusively by all of k random variables, without overcounting redundant information in attribute subsets. Thus, it finds irreducible and unexpected patterns in data that are necessary to learn from data [13].

We expect that exploiting this new source of information can help machine learning algorithms to improve their performance. For example, attribute interaction is helpful in domains where the lack of expert knowledge hinders the selection of very informative attributes sets by finding interacting attributes needed for learning. Another example is the case when the attribute representation is primitive and attribute relationships are more important than the attributes themselves. Then, similarity-based learning algorithms will probably fail, because the proximity in the instance space is not related to classification in this domain. Both cases apply to multimedia settings.

The k-way interaction information as found in [6] for a subset $\mathcal{S} \subseteq \mathcal{X}$ of all attributes $\mathcal{X} = \{X_1, X_2, ..., X_n\}$ is defined as:

$$\begin{aligned} I(\mathcal{S}) = & -\textstyle\sum_{\mathcal{T}\subseteq\mathcal{S}}(-1)^{|\mathcal{S}|-|\mathcal{T}|}H(\mathcal{T}) \\ & = I(\mathcal{S}\setminus X|X) - I(\mathcal{S}\setminus X), X \in \mathcal{S} \end{aligned} \quad (1)$$

with the entropy defined as:

$$H(\mathcal{S}) = -\sum_{\bar{v}\in\bar{\mathcal{S}}} P(\bar{v}) log_2 P(\bar{v}), \quad (2)$$

where the cartesian product of the sets of attribute values $\bar{\mathcal{X}} = \mathcal{X}_1 \times \mathcal{X}_2 \times ... \times \mathcal{X}_n$ is used. The feature interaction for $k = 1$ reduces to the single entropy, for $k = 2$ to the well known mutual information and for $k = 3$ attributes to McGill's multiple mutual information:

$$I(A;B) = H(A) + H(B) - H(A,B) \quad (3)$$

$$\begin{aligned} I(A;B;C) = & \quad I(A;B|C) - I(A;B) \\ = & \quad H(A,B) + H(A,C) + H(B,C) \\ & - H(A) - H(B) - H(C) - H(A,B,C). \end{aligned} \quad (4)$$

According to this definition, 3-way information interaction will be only zero iff A and B are conditionally independent in the context of C, because then $I(A;B|C) = I(A;B)$. So it gives only the information exclusively shared by all three attributes.

Information interactions as defined here are stable and unambiguous, since adding new attributes does not change already existing interactions, but only adds new ones. Furthermore they are symmetric and undirected between attribute sets. In general, two levels of interactions are important: (1) relevant non-linearities between the input attributes $I(A;B;C)$, which can be used in unsupervised learning and (2) interactions $I(A;B|L)$ between the input attributes A, B and their indicator or class label L, which are needed in supervised learning.

Information interaction is not to be confound with multi-information as presented in [11]. This dependence measure is based on the Kullback-Leibler divergence between the joint probability of $X_i, i = 1...M$ attributes and their marginals:

$$I_{\text{multi}}(X) = \sum_i H(X_i) - H(X) = \sum_x P(x) log_2 \frac{P(x)}{\prod_i P(x_i)}$$

Multi-information also results for $i = 2$ in mutual information, but for $i = 3$ attributes, it differs from the information interaction, because it results in:

$$I_{\text{multi}}(A,B,C) = H(A) + H(B) + H(C) - H(A,B,C).$$

Hence, it can capture higher order statistical dependencies, but is not taking into account the context of the other variables, the pairwise interactions. This way, multi-information overfits the k-way mutual information by counting redundant feature dependencies several times.

An important characteristic of feature information interactions is that it can result in positive and negative values. To information theoreticians, negative mutual information has no meaning. In [6], [7], an interpretation towards differentiating different basic types of feature interactions is given and is presented in the following subsections in more detail.

A. Positive interaction: Synergy

In case of positive interaction, it can be said that the process benefits from an unexpected synergy within the data. In statistics, this phenomena is also known as *moderating effect* and is discussed since a long time. Synergy occurs when A and B are statistically independent, but get dependent in the context of C as is symboled in figure 1. In the ordinary graph of the variables and their feature interactions, this is depicted by small pairwise relations between $I(A,C)$, $I(B,C)$ and $I(A,B)$ which are in sum smaller than the 3-way interaction $I(A,B,C)$. In [6], this type of interaction is described as observational, because the relationships between the features can only be found by looking at all of them at once.

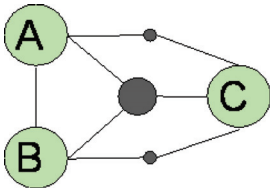

Fig. 1. Synergy: positive 3-way feature interaction with A, B being independent

Myopic feature selection strategies are unable to exploit the synergy in the data. Synergetic feature subsets can be exploited by feature selection and construction, as is shown in Section V.

B. Negative interaction: Redundancy

Negative interaction occurs when attributes partly contribute redundant information in the context of another attribute, which leads to a reduction of the overall dependence. It is visualized in Figure 2 where the redundant attributes A, B are related to a third attribute or class label C. The graph clarifies that if the sum of pairwise interactions $I(A,C)$, $I(B,C)$ and $I(A,B)$ includes the high interaction of a correlated pair, the information interaction or gain $I(A,B,C)$ drops to a negative value representing the information loss by redundancy. This type of interaction is called representational, because it includes some conditions on all involved attributes.

The negative influence of redundancy can theoretically be resolved by eliminating unneeded attributes, but in practice, it could be also advantageous in the case of noisy data.

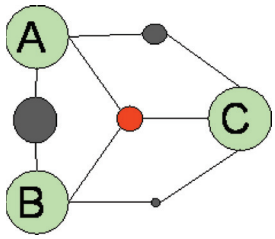

Fig. 2. Redundancy: negative 3-way feature interaction with A, B being dependent

In all cases, myopic voting functions that are based on the independence assumption as well as other fusion algorithms that use only local dependencies are confused by positive and negative feature interaction, which results in decreased performance of information fusion systems. Furthermore, experiments show that it is generally harder to resolve the influence of negative interactions.

IV. Experiments on artificial data

To validate our model, we first present extensive test results on artificial data that analyze how feature relationships can be determined with feature information interaction and multi information. The goal is to understand why these dependence measures are superior to traditional pairwise measures like correlation and mutual information. These findings can then be used to better exploit relevant feature dependencies in real world data by learning complex feature relationships and hence improving for example information fusion in multimedia problems (see section V).

A. Simple tasks: AND and OR

First, we investigate the two basic relations that can interconnect features: the boolean AND and OR. The test sets were created as follows with 10'000 samples:

- AND: The domain consists of six random and hence independent informative binary attributes. The class label is true if $And(s) = [s_1 \bigwedge s_2 \bigwedge s_3]$, which results in a set with circa 15% true class labels. Disregarding the label, it is a completely random set, thus no feature relations can be found. Otherwise, it is an easy task and should be solved by all dependence measures.
- AND 25 and 50: These sets are created by the AND rule such that they hold 25% and 50% true class labels, respectively. All missing values are filled in randomly. These sets also contain patterns that can be found without regarding the class label.
- OR: The domain has again six independent attributes where $Or(s) = [s_1 \bigvee s_2 \bigvee s_3]$. Here the set has circa 50% true class labels.
- OR 25: That is why we added just one rule-based set having 25% true class label, which is created in the same way as before the additional AND sets.

For each of the 5 sets, we calculated the dependence measures, absolute correlation, mutual information, 3-way feature interaction, 3-way multi information, 4-way feature interaction and 4-way multi information, on the unsupervised input (the input conditioned on the class label) and the supervised input (where we use the class label as an input variable). The results are given in Table 1. We give the maximum dependency values (between brackets) and the associated feature numbers for the non-negative and pairwise measures as absolute correlation, mutual information and the multi-information. For the k-way feature information interaction, we give the minimum and maximum values, if they exist, with their associated feature numbers. For this first two tasks the features $1, 2, 3$ are dependent on the class label and all should be found by the measures.

As is clear from the data creation, the unsupervised input task for the AND and OR task are unsolvable. The AND, AND 25 and AND 50 task is solved easily by nearly all dependence measures. This is because even though the problem has three dependent attributes, it can be divided in pairwise sub-problems e.g. the relation between two dependent attributes or the relation between one dependent attribute and the class label and thus is solvable by pairwise dependence measures.

Regarding synergy and redundancy, we see that the 3-way feature information interaction determines most of its results with negative values which indicates redundancy. For 4-way information interactions, the result is less clear: the unsupervised task is never solved, the task with conditioned class labels results in high synergies (positive values) and for the class label and feature input it results in redundancy. For the latter, it should be noted that those negative result values for finding the class features are in fact the maximum values of the dependence measure distribution. In the minimum and therewith high redundant values, the features being not dependent on the class can be found. The theoretically weaker multi information outperforms one time the feature interactions in the AND 25 unsupervised case.

The OR and OR 25 tasks are more difficult to solve, because the problem is not dividable into subproblems, but all of the dependent attributes have to be taken into account at once. That is why all pairwise measures fail. The 3-way feature interactions finds in the dependent features synergies, the 4-way feature interactions for the OR tasks result in similar values as for the AND tasks except that there are no synergies for the class conditioned case, but negative values which are again the maxima of the dependence measure histogram.

It can be concluded from this experiment that 3-way feature information interaction determines redundancy for AND connected attributes and synergy for OR connected ones. In 4-way feature information interaction, the class-dependent attributes can be found by taking the maximum value and the class-independent attributes by regarding the minima. Even though multi information finds more often the class dependent attributes correctly it has the disadvantage of not showing the synergy and redundancy informations. Pairwise dependence measures can only find redundant features in AND connections.

B. complex and hierarchical boolean constructs

In this experiment, we investigate if the dependence measures can resolve more complex feature relations like for example in feature hierarchies and the XOR problem. The following data sets have been created with 10'000 samples:

- ParityAnd(i,j): Domain with 12 random attributes where the class label is true if $oddand(s) = odd(s_i, s_6) \bigwedge odd(s_7, s_j)$. Whereas $odd(s)$ is true the sequence has an odd number of ones [12].
- ParityXOR(i,j): Domain with 12 random attributes where the class label is true if $oddxor(s) = NOT(odd(s_i, s_7)) \bigvee odd(s_8, s_j)$. Whereas $odd(s)$ is true the sequence has an odd number of ones [12].

The results for each task and dependence measure are shown in Table 2. We dropped from the table the unsupervised case since it is, by definition, completely random and hence contains no information.

At first sight the ParityAnd tasks $5-8$ and $4-8$ are of order $N = 4$ and $N = 5$ respectively. But they can also be divided in two sub-problems that are AND-connected and of size $N = 2$ or $N = 3$. They are (at least for the class conditioned case) solvable by the pairwise dependence measures, but due to the hierarchy, they fail for the supervised case. For the 3-way and 4-way feature interactions and multi-information, the relation hierarchy is robustly resolved. With the latter, in the class-conditioned case and for the ParityAnd $5-8$, even all dependent attributes are found in combination. In the results of the ParityAnd $4-8$, the feature information interactions are the first time outperforming the multi-information, because in two cases it finds both dependent subsets, whereas the multi-information resolves only the $N = 2$ sub-problem. The Parity XOR task behaves similar to the $N = 4$ ParityAnd task. This experiment showed that feature information interaction outperforms multi-information for more complex attribute relationships.

V. Classification experiments

In this section, we now present classification experiments that compare our approach of feature selection and construction based on feature information interaction to a baseline system that uses no feature selection and systems that do a feature selection based on correlation and mutual information. The goal is to determine if the knowledge of synergies and redundancies can improve a classification task and how it is best exploited.

As test collection, we chose the Washington collection, which consists of 886 images that are annotated with 1 to 10 keywords and are grouped into 20 classes. The feature set that we extracted from the raw data consists of the global color and texture histograms which have 166 and 165 features respectively. The text component is represented by the term

Table 1. Results for AND and OR tasks with each row being: $D(..)$ unsupervised input, $D(..|L)$ input conditioned on class label and $D(.., L)$ supervised input by joining attributes and class label

AND		absolute correlation	2D mutual information	3D feature interaction	3D multi information	4D feature interaction	4D multi information
	$D(..)$	–	–	–	$2, 3(1.0)$	$2, 3, 6, 3(-0.002)$ $6, 3, 5, 4(-2.57)$	$2, 2, 3, 6(1.08)$
	$D(..\|L)$	$\mathbf{1}, \mathbf{2}(0.4)$ $\mathbf{2}, \mathbf{3}(0.4)$	$\mathbf{1}, \mathbf{2}(1.8)$ $\mathbf{2}, \mathbf{3}(1.8)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(-2.9)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(2.9)$	$\mathbf{1}, \mathbf{1}, \mathbf{2}, \mathbf{3}(3.99)$ $6, 4, 5, 5(-5.18)$	$\mathbf{1}, \mathbf{1}, \mathbf{2}, \mathbf{3}(3.99)$
	$D(.., L)$	–	$\mathbf{1}(0.1)$ $\mathbf{2}(0.1)$ $\mathbf{3}(0.1)$	$\mathbf{1}, \mathbf{2}(0.019)$ $\mathbf{2}, \mathbf{3}(0.021)$	$\mathbf{1}, \mathbf{2}(0.29)$ $\mathbf{1}, \mathbf{3}(0.28)$	$\mathbf{2}, \mathbf{1}, \mathbf{3}(-4.56)$ $6, 5, 4(-5.18)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(0.53)$
AND 25	$D(..)$	$\mathbf{1}, \mathbf{2}(0.2)$ $\mathbf{2}, \mathbf{3}(0.2)$	–	$\mathbf{1}, \mathbf{2}, \mathbf{3}(0.005)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(0.093)$	$2, 3, 6, 3(0.12)$ $6, 5, 3, 4(-2.023)$	$\mathbf{1}, \mathbf{1}, \mathbf{2}, \mathbf{3}(1.05)$
	$D(..\|L)$	–	$\mathbf{1}, \mathbf{2}(1.8)$ $\mathbf{2}, \mathbf{3}(1.8)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(-2.86)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(2.86)$	$\mathbf{1}, \mathbf{1}, \mathbf{2}, \mathbf{3}(3.82)$ $6, 5, 6, 4(-5.12)$	$\mathbf{1}, \mathbf{1}, \mathbf{2}, \mathbf{3}(3.87)$
	$D(.., L)$	$\mathbf{1}(0.5)$ $\mathbf{2}(0.5)$ $\mathbf{3}(0.5)$	$\mathbf{1}(0.2)$ $\mathbf{2}(0.2)$ $\mathbf{3}(0.2)$	$\mathbf{1}, \mathbf{2}(-0.029)$ $\mathbf{1}, \mathbf{3}(-0.033)$	$\mathbf{1}, \mathbf{2}(0.41)$ $\mathbf{1}, \mathbf{3}(0.41)$	$4, 5, 6(-5.62)$ $\mathbf{1}, \mathbf{2}, \mathbf{3}(-4.67)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(0.61)$
AND 50	$D(..)$	$\mathbf{1}, \mathbf{2}(0.3)$ $\mathbf{2}, \mathbf{3}(0.3)$	$\mathbf{1}, \mathbf{2}(0.1)$ $\mathbf{2}, \mathbf{3}(0.1)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(-0.008)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(0.21)$	$1, 3, 6, 1(0.41)$ $3, 4, 3, 5(-2.71)$	$1, 2, 5, 5(1.08)$
	$D(..\|L)$	–	$\mathbf{1}, \mathbf{2}(1.6)$ $\mathbf{2}, \mathbf{3}(1.6)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(-2.44)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(2.44)$	$\mathbf{1}, \mathbf{1}, \mathbf{2}, \mathbf{3}(3.27)$ $6, 5, 6, 4(-5, 36)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{3}(3.26)$
	$D(.., L)$	$\mathbf{1}(0.6)$ $\mathbf{2}(0.6)$ $\mathbf{3}(0.6)$	$\mathbf{1}(0.3)$ $\mathbf{2}(0.3)$ $\mathbf{3}(0.3)$	$\mathbf{1}, \mathbf{2}(-0.078)$ $\mathbf{1}, \mathbf{3}(-0.070)$	$\mathbf{1}, \mathbf{2}(0.62)$ $\mathbf{1}, \mathbf{3}(0.61)$	$4, 5, 6(-5.99)$ $\mathbf{1}, \mathbf{2}, \mathbf{3}(-4.09)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(0.92)$
OR	$D(..)$	–	–	–	–	$1, 2, 1, 6(-2.5)$	$1, 1, 5, 6(1.0)$
	$D(..\|L)$	–	–	$\mathbf{1}, \mathbf{2}, \mathbf{3}(0.99)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(1.0)$	$\mathbf{1}, \mathbf{2}, \mathbf{2}, \mathbf{3}(-3.50)$ $5, 6, 4, 3(-5.99)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{3}(2.0)$
	$D(.., L)$	–	–	–	–	$\mathbf{1}, \mathbf{2}, \mathbf{3}(-4.99)$ $4, 3, 6(-5.99)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(1.0)$
OR 25	$D(..)$	–	–	$\mathbf{1}, \mathbf{2}, \mathbf{3}(0.052)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(0.052)$	$4, 3, 4, 1(-2.51)$	$\mathbf{1}, \mathbf{1}, \mathbf{2}, \mathbf{3}(1.05)$
	$D(..\|L)$	–	–	$\mathbf{1}, \mathbf{2}, \mathbf{3}(0.99)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(1.0)$	$\mathbf{3}, \mathbf{1}, \mathbf{1}, \mathbf{2}(-3.49)$ $4, 6, 2, 3(-5.99)$	$\mathbf{1}, \mathbf{1}, \mathbf{2}, \mathbf{3}(2.0)$
	$D(.., L)$	–	–	–	–	$\mathbf{1}, \mathbf{2}, \mathbf{3}(-5.43)$ $6, 5, 4(-5.67)$	$\mathbf{1}, \mathbf{2}, \mathbf{3}(0.25)$

frequencies of each image keywords, where the dictionary size is 295.

The classification is done in all experiments with the support vector machine (SVM) algorithm using a RBF kernel following two different fusion strategies: (1) early information fusion (data fusion) that applies one SVM to the concatenated feature vector and (2) late information fusion (hierarchical fusion), where first a SVM is applied to each modality (color, texture, text), then these results are combined with another SVM. For both cases, we ran a cross-validation to optimize the parameters of the SVM. As training set, we randomly selected for each run and class 7 positive and 7 negative examples, the rest was used as test set. The experiments were ran with the one-againt-all classification strategy, where the precision values were averaged over all classes. Finally, the presented results are also averaged over ten runs, if nothing is otherwise stated.

A. Feature selection

A simple but nevertheless common strategy for feature selection approaches that exploit statistical dependencies in a supervised setting is the calculation of the pairwise relationships between the attributes and the class labels [3]. As dependency measures to compare our approach with, we chose absolute correlation and mutual information. After ranking the attributes in descending order, the best N are selected as input for the subsequent classification, where we set $N = 70$ for the following experiments.

Our feature selection strategy based on feature information interactions works similar. First, we calculate empirically all the 3-way feature interaction $I(A, B, L)$ between two features A, B and the class label L. Then again, the features are ranked according to their involvement in a 3-way interaction, where we discern three different strategies: (1) absolute interactions, (2) positive values (synergy) and (3) negative values (redun-

Table 2. complex and hierarchical problems

PartiyAnd $5-8$		absolute correlation	2D mutual information	3D feature interaction	3D multi information	4D feature interaction	4D multi information
	$D(..\|L)$	$\mathbf{5},\mathbf{6}(0.1)$ $\mathbf{7},\mathbf{8}(0.1)$	$\mathbf{5},\mathbf{6}(1.0)$ $\mathbf{7},\mathbf{8}(1.0)$	$\mathbf{5},\mathbf{5},\mathbf{6}(-1.0)$ $\mathbf{7},\mathbf{8},\mathbf{8}(-1.0)$	$\mathbf{5},\mathbf{5},\mathbf{6}(2.0)$ $\mathbf{7},\mathbf{8},\mathbf{8}(2.0)$	$\mathbf{7},\mathbf{8},\mathbf{5},\mathbf{6}(-3.49)$ $4,3,2,9(-5.99)$	$\mathbf{5},\mathbf{6},\mathbf{7},\mathbf{8}(2.0)$
	$D(..,L)$	–	–	$\mathbf{5},\mathbf{6}(0.31)$ $\mathbf{7},\mathbf{8}(0.32)$	$\mathbf{5},\mathbf{6}(0.31)$ $\mathbf{8},\mathbf{7}(0.32)$	$\mathbf{5},\mathbf{5},\mathbf{6}(-3.65)$ $\mathbf{7},\mathbf{7},\mathbf{8}(-3.66)$ $4,11,10(-5.67)$	$\mathbf{5},\mathbf{5},\mathbf{6}(1.31)$ $\mathbf{7},\mathbf{7},\mathbf{8}(1.32)$
ParityAnd $4-8$	$D(..\|L)$	$\mathbf{7},\mathbf{8}(0.1)$	$\mathbf{7},\mathbf{8}(1.0)$	$\mathbf{4},\mathbf{5},\mathbf{6}(0.99)$ $\mathbf{7},\mathbf{7},\mathbf{8}(-1.0)$	$\mathbf{8},\mathbf{7},4(1.0)$ $\mathbf{7},\mathbf{7},\mathbf{8}(2.0)$	$\mathbf{7},\mathbf{7},\mathbf{7},\mathbf{8}(0.001)$ $5,3,1,2(-5.99)$ $4,\mathbf{7},\mathbf{7},\mathbf{8}(-1.02)$	$\mathbf{7},\mathbf{7},\mathbf{7},\mathbf{8}(3.0)$
	$D(..,L)$	–	–	$\mathbf{7},\mathbf{8}(0.31)$	$\mathbf{7},\mathbf{8}(0.31)$	$\mathbf{7},\mathbf{8},\mathbf{7}(-3.66)$ $\mathbf{4},\mathbf{5},\mathbf{6}(-5.36)$ $12,11,9(-5.67)$	$\mathbf{7},\mathbf{7},\mathbf{8}(1.31)$
ParityXOR $6-9$	$D(..\|L)$	$\mathbf{6},\mathbf{7}(0.3)$ $\mathbf{8},\mathbf{9}(0.3)$	$\mathbf{6},\mathbf{7}(0.1)$ $\mathbf{8},\mathbf{9}(0.1)$	$\mathbf{6},\mathbf{7},\mathbf{7}(-0.09)$ $\mathbf{8},\mathbf{8},\mathbf{9}(-0.08)$	$\mathbf{7},\mathbf{7},\mathbf{6}(1.09)$	$\mathbf{6},\mathbf{7},\mathbf{8},\mathbf{9}(-5.53)$	$\mathbf{6},\mathbf{7},\mathbf{8},\mathbf{9}(0.41)$
	$D(..,L)$	–	–	$\mathbf{6},\mathbf{7}(0.32)$ $\mathbf{8},\mathbf{9}(0.31)$	$\mathbf{7},\mathbf{6}(0.32)$ $\mathbf{8},\mathbf{9}(0.31)$	$\mathbf{7},\mathbf{7},\mathbf{6}(-3.65)$ $\mathbf{8},\mathbf{8},\mathbf{9}(-3.65)$	$\mathbf{7},\mathbf{6},\mathbf{6}(1.31)$ $\mathbf{8},\mathbf{9},\mathbf{9}(1.31)$

dancy). Again the first 70 attributes are selected.

The full calculation of the 3-way feature information interaction matrix is not feasible even for a rather small collection like Washington, because its size grows exponentially with the size of the feature vector. That is why we applied random subsampling from a normal distribution in the feature and class label spaces. The sample size was set to $80'000$, which largely under samples the space of size $[626, 626, 20]$. One should keep in mind therefore that all the following results are based on an incomplete interaction matrix.

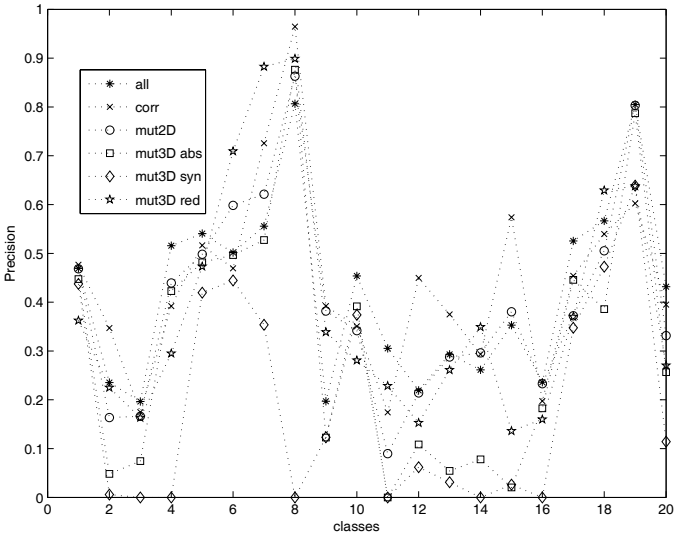

Fig. 3. Precision without and with different feature selection algorithms (hierarchical classification)

The mean average precision over the runs and classes of the Washington collection are given in Table 3 for the hierarchical (`hier`) and data fusion strategy (`data`). Additionally we list the number of times an approach is performing best for the 20 classes (best), which can be more than one approach per class. The algorithms tested are the baseline system with all features (all), the feature selection based on absolute correlation (corr), mutual information (mut2D), absolute 3-way feature interaction (abs(mut3D)), synergy (syn) and redundancy (red). Figures 3 and 4 depict the precisions for each class.

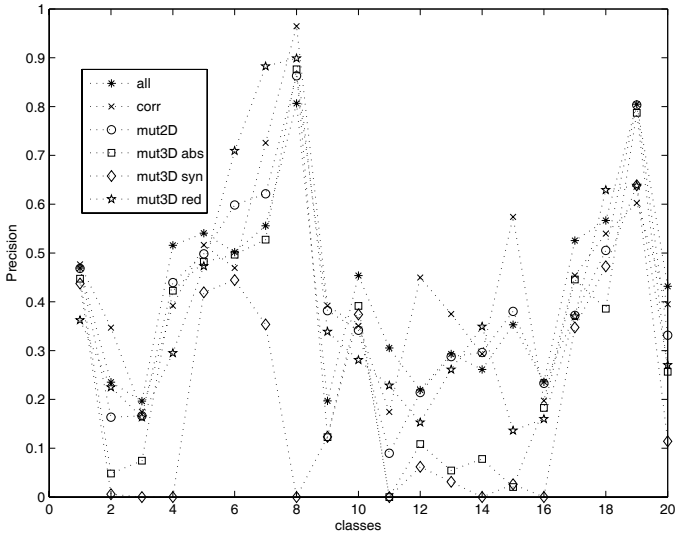

Fig. 4. Precision without and with different feature selection algorithms (data classification)

In average, the correlation-based feature selection performs best in the hierarchical fusion as well as in the data fusion, whereas the latter is strongly outperformed by the hierarchical fusion approach. If we count the number best performing pre-

Table 3. Mean average precision and best value counts for classification on the full feature set and the tested feature selection methods

	all	corr	mut2D	abs(m3D)	syn	red
hier	0.423	**0.443**	0.403	0.310	0.193	0.391
best	**10**	8	3	–	–	4
data	0.140	**0.330**	0.191	0.088	0.085	0.100
best	3	**18**	3	2	2	2

cision values the classification system without feature selection is best in case of hierarchical fusion. There it is also in terms of mean precision not significantly inferior to the correlation-based system. The place of the third best system is shared by the feature selection based on mutual information and the one based on redundancy. The first has a better average precision value, but the second has the best value for more classes. The systems based on absolute information interactions and synergetic features are completely unfeasible.

We conclude from this experiment that the baseline with a hierarchical fusion over the modalities is still a hard baseline to improve. The standard feature selection approach based on correlation has not improved it significantly. Concerning the performance of our approaches based on information interaction, we conclude that with feature selection only the redundancy information can be exploited, but this is not sufficient to attain an equal performance as the baseline or the correlation based feature selection.

B. Feature construction

Synergies cannot be exploited by simple feature selection as we have seen in the previous experiment. So we apply a simple feature construction that is based on synergetic feature pairs found in the input data.

To do so, we calculated empirically all the feature interactions $I(A, B|L)$ between two features A, B conditioned on the class label L. Again the complete calculation is infeasible, so we under sampled the result matrix as described above. Then, we ranked the feature pairs according to their magnitude of interaction, where the highest interaction value represents a highly synergetic feature pair. Thus, the ranking gives us the order of relevance of the feature pairs in solving the classification task. In the runs, the best N are chosen as input. The feature construction itself is set up as a hierarchical SVM. First, we create a mid-level feature over each feature pair by using a SVM, which are then fused in a second level SVM towards the complete classification result.

The results for class 19 of the Washington collection are shown in Figure 5 for one to ten feature pairs. The hierarchical fusion on the full feature set (all) is outperformed by the feature construction based on synergetic features (fconstsyn) on four and more feature pairs. The redundant feature pairs (fconstred) perform much worse, hence redundancy cannot be exploited by feature construction.

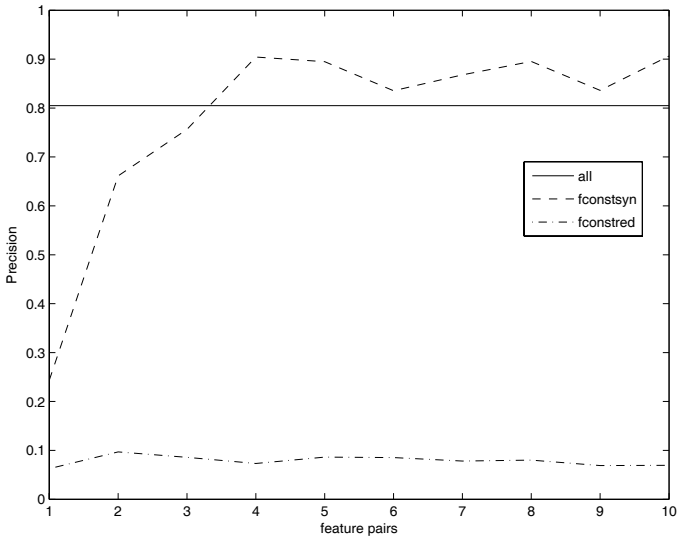

Fig. 5. Precisions for the full set compared to feature construction based on synergetic and redundant feature pairs (class = 19)

Our simple feature construction based on synergetic feature pairs can significantly outperform the baseline with only a small number of features used as input.

VI. Conclusion

We reviewed strategies for information fusion in high-dimensional feature spaces. They are classically based on the definition of mutual information and cannot detect information beyond pairwise relationships.

We then presented a promising new approach for fusing multimodal information which is of importance e.g. for multimedia document classification as well as many other applications. Experiments based on artificial data validated our approach. The algorithm based on fusing synergetic feature pairs is significantly outperforming the simple hierarchical fusion on the full feature set as well as the fusion based on greedy pairwise feature selectors, which are often used in machine learning. In direct comparison the feature information interactions perform best, but probably for problems consisting only of simple attribute relationships the multi information is sufficient and having the advantages of cheaper calculation and higher robustness.

The calculation of the feature information interactions still requires improvments. In turn, this may improve the actual results even further. A more profound understanding of synergy and redundancy and how it is represented in k-way feature information interaction is also needed. On a more general level the aim is to find a theoretic explanation of why interaction-based approaches outperform the myopic ones.

Acknowledgment

This work is funded by the European Project MultiMATCH (EU-IST-STREP#033104) and the Swiss NCCR (IM)2.

REFERENCES

[1] K. Barnard, P. Duygulu, and D. Forsyth, "Recognition as translating images into text," in *Internet Imaging IX, Electronic Imaging*, 2003. [Online]. Available: http://citeseer.ist.psu.edu/689616.html

[2] A. Bell, "The co-information lattice," in *4th Int. Symposium on Independent Component Analysis and blind Signal Seperation (ICA2003)*, 2003, pp. 921–926.

[3] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Machine Learning Research*, vol. 5, pp. 1531–1555, 2004.

[4] A. A. Freitas, "Understanding the crucial role of attribute interaction in data mining," *Artificial Intelligence Review*, vol. 16(3), pp. 177–199, 2001.

[5] B. Hanczar, J.-D. Zucker, C. Henegar, and L. Saitta, "Feature construction from synergic pairs to improve micro-array classification," vol. 23(21), pp. 2866–2872, 2007.

[6] A. Jakulin and I. Bratko, "Quantifying and visualizing attribute interactions," 2003.

[7] A. Jakulin and I. Bratko, "Analyzing attribute dependencies," in *Principles of Knowledge Discovery in Data (PKDD)*, 2003, pp. 229–240.

[8] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97(1-2), pp. 273–324, 1997.

[9] M. Kokar, J. Weyman, and J. Tomasik, "Formalizing classes of information fusion systems," *Information Fusion*, vol. 5, pp. 189–202, 2004.

[10] I. Kononenko, E. Simec, and M. Robnik-Sikonja, "Overcoming the myopia of inductive learning algorithms with relieff," *Applied Intelligence*, vol. 7(1), pp. 39–55, 1997.

[11] I. Nemenman, *Information theory, multivariate dependence and genetic networks*, eprint arXiv:q-bio/0406015, ARXIV, 2004.

[12] E. Perez, L.A. Rendell *Using Multidimensional Projections to Find Relations*, Proc. 12th International Conference on Machine Learning, 1995.

[13] I. Perez, "Learning in presence of complex attribute interactions: An approach based on relational operators," Ph.D. dissertation, 1997.

[14] N. Slonim, G. Atwal, G. Tkacik, and W. Bialek, "Estimating mutual information and multi-information in large networks," 2005.

[15] T. Westerveld and R. van Zwol, "Multimedia retrieval at inex 2006," in *ACM SIGIR Forum*, vol. 41(1), 2007, pp. 58–63.