Can Feature Information Interaction help for Information Fusion in Multimedia Problems?

Jana Kludas[1], Eric Bruno[1], and Stephane Marchand-Maillet[1]

University of Geneva, Switzerland,
jana.kludas@cui.unige.ch,
http://viper.unige.ch/

Abstract. The article presents the information-theoretic based feature information interaction, a measure that can describe complex feature dependencies in multivariate settings. According to the theoretical development, feature interactions are more accurate than current, bivariate dependence measures due to their stable and unambiguous definition. In experiments with artificial and real data we compare the empirical estimates of correlation, mutual information and 3-way feature interaction. We can conclude that feature interactions give a more detailed and accurate description of data structures that should be exploited for information fusion in multimedia problems.

1 Introduction

With the rise of the Web 2.0 and its tendency to populate the WWW more and more with images and videos multimedia related topics became lively discussed fields of research. In its core there is an essential need for information fusion due to the multi modal nature of multimedia data. Hence the fusion of multi modal data (e.g. text and images) has a large impact on algorithms like multimedia indexing, retrieval and classification, object recognition as well as for data preprocessing like feature selection or data model development. Information fusion has established itself as an independent research area over the last decades, but a general theoretic framework to describe general information fusion systems is still missing [6]. Up to today the understanding of how fusion works and by what it is influenced is limited. Probably that is one reason why in multimedia document retrieval for web applications the visual component is up to today lacking behind expectations as can be seen for example in the INEX 2006 [24] and 2007 Multimedia Track , where text-only based runs outperformed all others. As another example can be named the text-based image searches from Google, Yahoo! and others.

All work done so far on information fusion in multimedia settings can 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 statistical dependencies. Both approaches have been applied in multimedia processing problems equally successfully - for some

tasks the fusion on independent sources outperforms the algorithms based on dependent sources, on other tasks it is the other way around. Neither of these approaches is in general superior.

Aligned to the second approach we like to investigate another way of analyzing input data for multimedia problems based on feature information interactions with the long term goal of information fusion performance improvement. This multivariate, information theoretic based dependence measure is more accurate in finding the data's structure e.g. situations, where the independence assumption is sufficient and where the dependency between the input data is not negligible.

Information interaction is superior to traditional dependence measures due to its consistent definition, its global application to the whole feature set and its capture of linear and higher order statistical dependencies. Considering this new definition of feature interactions current machine learning algorithms do not treat the feature's statistical dependencies properly. Hence the investigation of feature interactions in multimedia data could help to improve the information fusion and hence the whole performance of the entailed retrieval and classification algorithm.

In Section 2 we discuss in more detail state-of-the-art fusion approaches with independent and dependent input data and their shortcomings. Thereafter we present in Section 3 the idea of feature interaction information and how it can help to improve information fusion algorithms. In Section 4 we give the results of data analysis experiments with artificial and real data, which is followed by the conclusions in Section 5.

2 Related Work

Our article discuss the problem of information fusion, but most of the related work can be found in multimedia processing where information fusion is only implicitly treated as one part of the problem. We review some example approaches and explain when and why they can fail.

In early years of information fusion research scientists fused different sources by assuming independence between them as in one of the first works on classifier and decision fusion on fusing neural network outputs [4]. The independence assumption is still widely used in machine learning as e.g. in the naive Bayes classifier. Its success is based on its simplicity in calculation and the learned models, as well as its robustness in estimating the evidence [18]. Approaches that fuse independent or complementary sources mostly belong to classifier and decision fusion, where first each modality of the input is treated separately and then a final decision is based on the single results. Applications that can be found in literature are for example multimedia retrieval [14, 12], multi modal object recognition [5], multi-biometrics [7] and video retrieval [15].

Despite its successful application for some problems it seems to fail completely for others. In [7] it is shown that the violation of the independence assumption hurts the information fusion performance. So a trade off between

simple and fast calculated results and their accuracy is necessary. That loss in performance was empirically undermined in [9], where they showed that the maximum performance improvement in a multi-biometrics application can be only achieved, if the statistical dependencies between the modalities are taken into account. Independence assumption based algorithms are also called myopic, because they treat all attributes as conditionally independent given the class label [25].

To circumvent the problem of attribute dependencies in data other approaches try to create independence with the help of linear transformation methods like principal and independent component analysis (PCA/ICA), factor analysis and projection pursuit as reviewed in [19]. Unfortunately these methods are not sufficient to eliminate all dependencies in the data, since they target only pairwise and linear feature dependencies [20]. In addition the authors showed empirically that their multi modal object recognition problem is affected by higher order dependency patterns. A similar result was found in [16]. In the multimedia classification task the Support Vector Machine (SVM) approach using an ICA-based feature selection was outperformed by a SVM on the original data set. Multimedia processing approaches that exploit explicitly attribute dependencies fuse the information preferably at data or feature level. Example applications are multimedia summarization [1], text and image categorization [3], multi modal image retrieval [13] and web document retrieval [8]. Those approaches exploit all some form of attribute dependency at data level like co-occurrence (LSI [28]), correlation (kCCA [16]) or mutual information. As examples for late fusion, where classifier dependencies are exploited, can be named copula functions [27] or nonlinear fusion algorithms based on SVM's [2].

The most important shortcoming of those algorithms is that they only take bivariate dependencies into account, even though they work in a multivariate setting [21]. High level feature relationships such as conditional dependencies of a feature pair to a third variable e.g. the class label are neglected. For now there exists no prove that this higher order dependencies have an impact on the performance of multimedia processing systems, but in [22] their exploitation led to a performance improvement.

3 Feature Information Interaction

Before the introduction of feature interaction by [17, 18] 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 [23] is based on McGill's multivariate generalization of Shannon's mutual information. It describes the information that is shared by all of k random variables, without overcounting redundant information in attribute subsets. So it finds irreducible and unexpected patterns in data that are necessary to learn from data [26].

This general view of attribute interactions could help machine learning algorithms to improve their performance. For example attribute interactions can be

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 fail, because the proximity in the instance space is not related to classification in this domain.

Two levels of interactions can be differentiated: (1) relevant non-linearities between the input attributes, which are useful in unsupervised learning and (2) interactions between the input attributes and the indicators or class labels, which is needed in supervised learning. The k-way interaction information as found in [17] for a subset $\mathcal{S}_i \subseteq \mathcal{X}$ of all attributes $\mathcal{X} = \{X_1, X_2, ..., X_n\}$ is defined as:

$$I(\mathcal{S}) = -\sum_{\mathcal{T} \subseteq \mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{T}|} H(\mathcal{T}) = I(S \setminus X|X) - I(\mathcal{S} \setminus X), X \in \mathcal{S} \tag{1}$$

with the entropy defined as:

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

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:

$$I(A; B) = H(A) + H(B) - H(A, B) \tag{3}$$

$$\begin{aligned} I(A; B; C) = & \quad I(A; B|C) - I(A; B) & (4)\\ = & \quad H(A, B) + H(A, C) + H(B, C) & (5)\\ & - H(A) - H(B) - H(C) - H(A, B, C). & (6) \end{aligned}$$

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 the involved attributes. Information interactions are stable and unambiguous, since adding new attributes is not changing already existing interactions, but only adding new ones. Furthermore they are symmetric and undirected between attribute sets.

It is not to be confused with multi-information as presented in [21]. This dependence measure is based on the Kullback-Leibler divergence between the joint probability of $X_i, i = 1...M$ attributes and their marginals:

$$I_{multi}(X) = \sum_i H(X_i) - H(X) = \sum_{x_i} P(x) log_2 \frac{P(x)}{\prod_i P(x_i)} \tag{7}$$

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

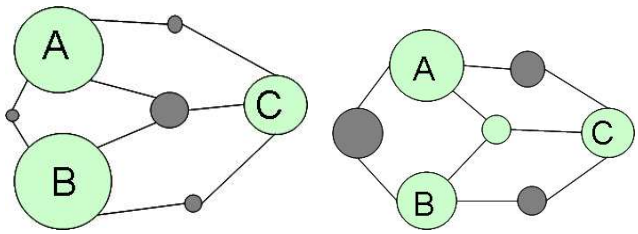
$$I_{multi}(A, B, C) = H(A) + H(B) + H(C) - H(A, B, C). \tag{8}$$

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

Another interesting point about feature information interaction is that it results in positive and negative values, which represent two different types of feature interactions. An explanation using synergy and redundancy between attributes that was given in [17, 18], is presented in the following.

3.1 Positive interaction: Synergy

In case of positive interactions the process benefits from an unexpected synergy in the data. In statistics this phenomena is called moderating effect and is known a long time. Synergy occurs when A and B are statistical independent, but get dependent in the context of C as can be seen in Figure 1(a). In [17] 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. Myopic feature selections are unable to exploit the synergy in the data. It can be exploited e.g. for feature selection in supervised learning or for feature construction in the unsupervised case.



(a) synergy (b) redundancy

Fig. 1. Interaction diagrams of different types of information interactions between A, B and C

3.2 Negative interaction: Redundancy

Negative interactions occur when attributes partly contribute redundant information in the context of another attribute, which leads to a reduction of the overall dependence. It is shown in Figure 1(b) on behalf of the redundant attributes A, B towards a third attribute C. This type of interaction is also called representational, because it includes some conditions on all involved attributes. In supervised learning the negative influence of redundancy can be resolved by eliminating unneeded redundant attributes, but it could be advantageous in unsupervised learning in the case of noisy data.

In any case myopic voting function that are based on the independence assumption as well as fusion algorithms that use only local dependencies are confused by positive and negative feature interactions, which results in decreased information fusion performance. In general it is harder to resolve the influence of negative interactions.

In the following section we compare empirical estimates of correlation, mutual information and 3-way feature information interaction for artificial and real multi modal data to draw conclusions about their usefulness as dependence measure in information fusion.

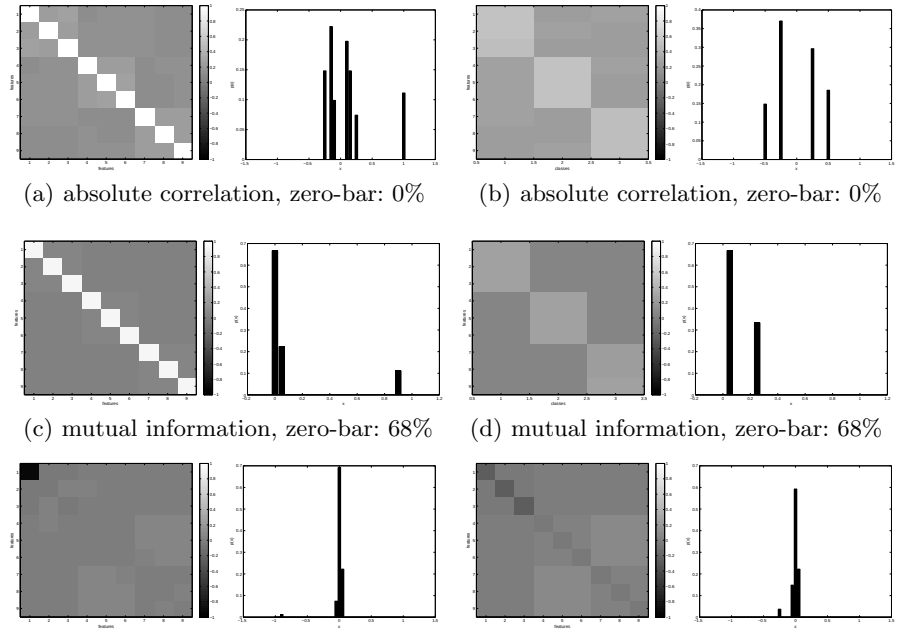
4 Experiments

For the objective evaluation of the different dependence measures we first conducted tests on simple artificial data sets, where the relations between the input variables as well as their relations towards the class labels are known.

The first artificial data set is based on an AND combination of 3 binary variables defining one of the 3 classes. Additional input variables are filled with random values. Hence the intra-class variables are dependent on each other and their class label, but independent to the other six input variables.

Figure 2 shows the empirical estimates and histograms of the correlation matrix, the mutual information and the 3-way information interaction respectively for the unsupervised (features towards features) and the supervised (features towards class labels) case. In the both all dependence measures succeed in finding the 3 dependent intra-class variables, but with differences in accuracy.

Correlation, for example, is constantly overestimating the dependencies, because it shows no independence for the inter-class variables. Furthermore the knowledge of positive or negative correlation seem of no use for information fusion, but only the absolute magnitudes. Mutual information performs similarly in accuracy as information interaction. So it finds the inter-class independence of the input variables as well as the dependence of the intra-class variables. Finally, information interaction is giving the most detailed information about the data's structure. For the intra-class variables it results in negative interaction, which indicates redundancy. The inter-class information interactions are mostly zero and surprisingly it shows positive interactions, hence synergy, between the blocks of intra-class variables, where we are not sure yet how to explain this.



(a) absolute correlation, zero-bar: 0% (b) absolute correlation, zero-bar: 0%

(c) mutual information, zero-bar: 68% (d) mutual information, zero-bar: 68%

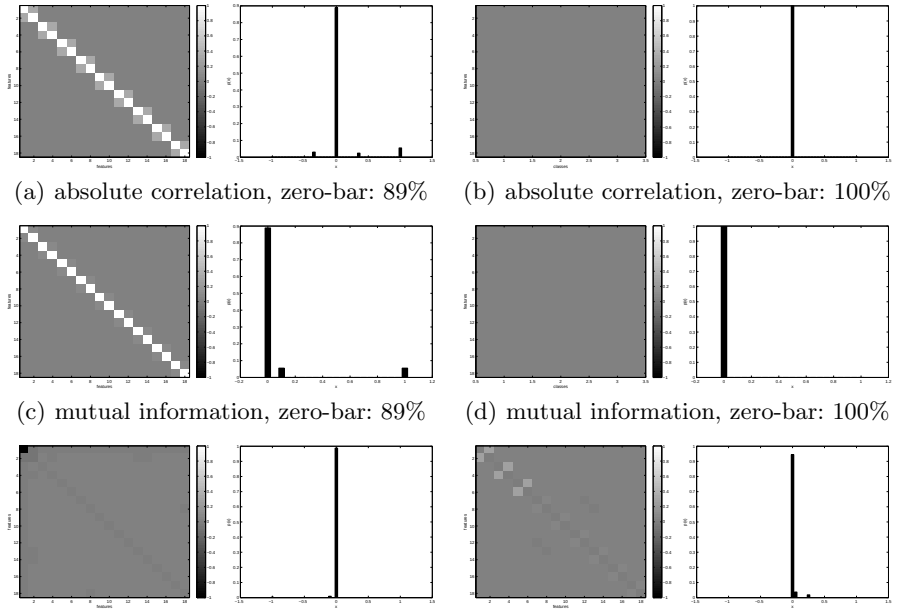
(e) 3-way interaction ($fv_3 = 1$) zero: 69% (f) 3-way interaction ($cl = 1$) zero: 60%

Fig. 2. Unsupervised (a,c,e) / supervised (b,d,f) case for AND combined artificial data

The second and more interesting artificial data set is based on the AND data set, but now each input variable is replaced by its XOR combination of 2 variables. Overall it has again 3 classes, where each depends now on 6 input variables. This new data set is a parity problem, which contains synergy between the XOR combined variables and their class labels.

Figures 3(a),3(c) and 3(e) show the empirical estimates and the histograms for the unsupervised case. Correlation finds independence between all variables except between the parity variables, where it results randomly in positive or negative correlations. Mutual information as well as the 3-way information interaction results show also only the dependence between the parity variables. So none of the investigated dependence measures finds all features that one class depends on in the unsupervised setting. We hope to find this hidden dependencies by investigating higher order information interactions in future work.

The results of the supervised case, that are presented in the Figures 3(b),3(d) and 3(f), show a clear advantage of information interaction over the other two dependence measures. Correlation and mutual information do not succeed in finding even the parity variables, because they are based only on bivariate relationships. Whereas information interaction finds synergy between the parity variables and detects all dependent variables of a class. As in the unsupervised case we hope to find the intuitively expected redundancies between the pairs of parity variables by regarding higher order information interactions.



(a) absolute correlation, zero-bar: 89% (b) absolute correlation, zero-bar: 100%

(c) mutual information, zero-bar: 89% (d) mutual information, zero-bar: 100%

(e) 3-way interaction ($fv_3 = 1$) zero: 99% (f) 3-way interaction ($cl = 1$) zero: 94%

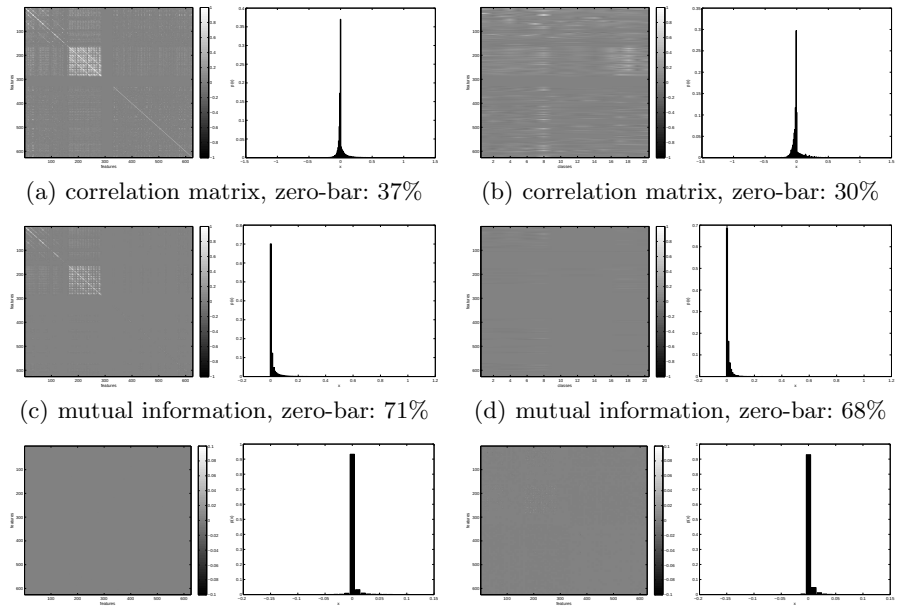
Fig. 3. Unsupervised (a,c,e) / supervised (b,d,f) case for OR combined artificial data

To summarize, it can be said that feature information interactions more accurately describe complex dependence structures in data sets by giving their irreducible patterns. This is especially true for parity problems. Furthermore it allows to differentiate feature relationships into synergies and redundancies, which we feel is useful knowledge to exploit in information fusion systems.

For the real data experiments we used the Washington collection, which consists of 886 images annotated with 1 to 10 keywords and grouped into 20 classes. The extracted feature set consists of the global color and texture histogram which have 165 and 164 features respectively. Additionally we constructed of the term frequencies of the keywords a textual feature vector of size 297.

This simple setting is in fact too simple to succeed with a classification or retrieval task. Intuitively global visual features and a handful of keywords are insufficient to discriminate any class. So we expect low relationships between the features in both: the unsupervised case and the supervised.

Ignoring the class labels we first investigated the feature dependencies for the unsupervised setting. We calculated a sampled version of the 3-way information interaction, where each sample consists of $k = 3$ random features out of the whole set. Figures 4(a),4(c) and 4(e) give the empirical estimates of the dependence measures and their histograms. As expected the feature information interactions show only little dependence in the feature set. Be aware that



(a) correlation matrix, zero-bar: 37% (b) correlation matrix, zero-bar: 30%

(c) mutual information, zero-bar: 71% (d) mutual information, zero-bar: 68%

(e) 3-way interaction ($fv_3 = 1$) zero: 95% (f) 3-way interaction ($cl = 1$) zero: 93%

Fig. 4. Unsupervised (a,c,e) / supervised (b,d,f) case for the Washington collection

the interaction diagrams are scaled between $[-0.1, 0.1]$ compared to $[-1, 1]$ for correlation and mutual information. So it is clearly visible that the latter two, both 2-way dependence measures, indicate much higher relationships (in number and magnitude) between the features. Hence one can state that they also overestimate the feature's dependencies for real data sets.

The results for the supervised setting are shown in Figures 4(b),4(d) and 4(f). Again the scale of the information interaction diagrams is set to $[-0.1, 0.1]$ for visibility reasons. Here the correlation between the features and their class labels results in high dependencies that are neither supported by the mutual information nor the 3-way feature information interaction. Mutual information overestimates as well a little the dependencies.

Experiments that compare end-to-end classification or retrieval results based on different feature selection or construction algorithms in multimedia problems have still to be done in future work. Until then the usefulness of feature information interactions in information fusion stays empirically unproven, but reasonable given that complex feature relationships can be estimated reliably.

5 Conclusions and Future Work

The article reviews the formal theory and characteristics of feature information interaction, an information-theoretic dependence measure. Through its stable

and unambiguous definition of feature relationships it can more accurately determine dependencies, because e.g. redundant contributions to the overall relationships are not overcounted.

Interestingly, information interaction can have positive and negative values, whereas until now it is not completely clear how to consistently resolve the negative ones. Positive interactions are synergies, that should be exploited, for example, by complicating the data model and using the feature's joint evidence.

Experiments on artificial data, where the feature dependencies are known, undermine the theoretically claimed superior performance of information interactions over bivariate dependence measures like correlation and mutual information especially for parity problems. These findings in the controlled setting fit also the tests on the real data of the Washington collection. The final prove of usefulness of feature information interactions for information fusion in classification or retrieval has to be done in future work.

Other directions of research will be the utilization of more complex multimedia data as e.g. the Wikipedia collection and tests with more sophisticated features like moment-based visual features.

References

1. A. B. Benitez, S. F. Chang, *Multimedia knowledge integration, summarization and evaluation*, Workshop on Multimedia Data Mining, 2002, pp. 23-26.
2. E. Bruno, N. Moenne-Loccoz, S. Marchand-Maillet, *Design of multimodal dissimilarity spaces for retrieval of multimedia documents*, to appear in IEEE Transaction on Pattern Analysis and Machine Intelligence, 2008.
3. G. Chechik, N. Tishby, *Extracting relevant structures with side information*, Advances in Neural Information Processing Systems, **15**, 2003.
4. K. Tumer, J. Gosh, *Linear order statistics combiners for pattern classification*, Combining Artificial Neural Networks, 1999, 127–162.
5. L. Wu, P.R. Cohen, S.L. Oviatt, *From members to team to committee - a robust approach to gestural and multimodal recognition*, Transactions on Neural Networks, **13(4)**, 2002, 972 - 982.
6. M. M. Kokar, J. Weyman, J.A. Tomasik, *Formalizing classes of information fusion systems*, Information Fusion, **5**, 2004, 189–202.
7. N. Poh, S. Bengio, *How do correlation and variance of base-experts affect fusion in biometric authentication tasks?*, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 53, 2005, pp. 4384–4396.
8. R. Zhao, W. I. Grosky, *Narrowing the semantic gap - improved text-based web document retrieval using visual features*, IEEE Trans. on Multimedia, **4(2)**, 2002, 189–200.
9. S. C. Dass, A. K. Jain, K. Nandakumar, *A principled approach to score level fusion in multimodal biometric systems*, Proc. of Audio- and Video-based Biometric Person Authentication (AVBPA), 2005, pp. 1049–1058.
10. S. Wu, S. McClean, *Performance prediction of data fusion for information retrieval*, Information Processing and Management, **42**, 2006, 899–915.
11. D. M. Squire, W. Mller, H. Mller, and J. Raki, *Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights*

and relevance feedback, in the 10th Scandinavian Conference on Image Analysis (SCIA'99), (Kangerlussuaq, Greenland), 1999, pp. 143–149.
12. T. Kolenda, O. Winther, L.K. Hansen, J. Larsen, *Independent component analysis for understanding multimedia content*, Neural Networks for Signal Processing, 2002, 757– 766.
13. T. Westerveld, A. P. de Vries, *Multimedia retrieval using multiple examples*, In International Conference on Image and Video Retrieval (CIVR'04), 2004, 344-352.
14. Y. Wu, K. Chen-Chuan Chang, E. Y. Chang and J. R. Smith, *Optimal multimodal fusion for multimedia data analysis*, MULTIMEDIA '04: Proc. of the 12th annual ACM international conference on Multimedia, ACM Press, 2004, pp. 572–579.
15. R. Yan, A. G. Hauptmann, *The combination limit in multimedia retrieval*, MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia, ACM Press, 2003, pp. 339–342.
16. A. Vinokurow, D.R. Hardoon and J. Shawe-Taylor, *Learning the Semantics of Multimedia Content with Application to Web Image Retrieval and Classification*, in Proceedings of Fourth International Symposium on Independent Component Analysis and Blind Source Separation, Nara, Japan, 2003.
17. A. Jakulin, I. Bratko, *Quantifying and Visualizing Attribute Interactions*, ArXiv Computer Science e-prints, Provided by the Smithsonian/NASA Astrophysics Data System, 2003.
18. A. Jakulin, I. Bratko, *Analyzing Attribute Dependencies*, Proc. of Principles of Knowledge Discovery in Data (PKDD), 2838, 2003, 229–240.
19. A. Hyvarinen, E. Oja, *Independent Component Analysis: Algorithms and Applications*, Neural Networks, 2000, 13(4-5),pp. 411-430.
20. N. Vasconcelos, G. Carneiro, *What is the Role of Independence for Visual Recognition?*, European Conference on Computer Vision, Copenhagen, 2002, 297 - 311.
21. I. Nemenman, *Information theory, multivariate dependence and genetic networks*, eprint arXiv:q-bio/0406015, ARXIV, 2004.
22. M.J. Pazzani, *Searching for Dependencies in Bayes Classifiers*, 1996, Learning from Data: AI and Statistics, Springer Verlag.
23. A.J. Bell, *The Co-Information Lattice*, 4th Int. Symposium on Independent Component Analysis and blind Signal Seperation (ICA2003), 2003, pp. 921–926.
24. T. Westerveld and R. van Zwol, *Multimedia Retrieval at INEX 2006*, 2007, ACM SIGIR Forum, 41(1), pp. 58-63.
25. I. Kononenko, E. Simec and M. Robnik-Sikonja, *Overcoming the myopia of inductive learning algorithms with RELIEFF*, Applied Intelligence, 7(1), 1997, pp. 39-55, Springer Netherlands.
26. I. Perez, *Learning in presence of complex attribute interactions: An Approach Based on Relational Operators*, PhD dissertation, University of Illinois at Urbana-Champaign, 1997.
27. K. Jajuga and D. Papla, *Copula Functions in Model Based Clustering*, in Studies in Classification, Data Analysis, and Knowledge Organization, Part 15, 2006, Springer Berlin Heidelberg.
28. T. Liu, Z. Chen, B. Zhang, W. Ma and G. Wu, *Improving Text Classification using Local Latent Semantic Indexing*, Fourth IEEE International Conference on Data Mining (ICDM'04), pp. 162-169, 2004.