Multimedia Retrieval and Classification for Web Content

Jana Kludas
CUI - University of Geneva, 24 rue Général Dufour, 1211 Geneva 4, Switzerland
jana.kludas@cui.unige.ch

The population of the World Wide Web with media of all types such as texts, images, videos and audio files in recent years raised the attractiveness of multimedia retrieval. With our work on the influence of dependencies between modalities and features we investigate why these approaches still do not perform convincingly better than plain text search approaches when applied to large, noisy collections like web content, even though these approaches have more information at their hands. This article suggests that, due to the size and noise, the modality's dependencies necessary for efficient information fusion becomes small and hard to exploit. Preliminary experiments with two multi modal collections underpin this statement.

Multimedia Retrieval, Multimedia Classification, multi modal Information Fusion.

1. INTRODUCTION

Some years ago the content of the World Wide Web consisted mainly of text. Since then, also due to the development of the Web 2.0, it became populated more and more with the other media types: images, audio files and videos. This created the necessity for algorithms to search and browse this new media content, but also the idea of combining information of different modalities to improve the information retrieval and classification approaches.

Multimedia retrieval and classification has received a lot of interest recently, which includes the introduction of multimedia tracks in evaluation workshops like TRECVID, IMAGECLEF and INEX. For now the multi modal fusion approaches helped for sure to improve image retrieval by using the aligned texts or keywords, but they still do not perform convincingly better than plain text-based search [1], even though having more information at their hands. This is caused on the one hand by the lack of understanding of the relation between low-level features and their semantics [7]. This is already a serious problem in simple image retrieval (semantic gap) and is enforced during fusion with the images' accompanied keywords and texts. On the other hand, many questions are open in information fusion. For example how to fuse dependent sources efficiently and how to predict the performance improvement that can be achieved by fusing different modalities, sources or samples.

To shed more light on the problems in multimedia retrieval and classification for web content the next section will review currently successful information fusion approaches and how they fuse different modalities. Section 3 explains then why current information fusion approaches probably have problems, when they are applied to noisy web content. The last section presents research directions and ideas for approaching a solution.

2. REVIEW OF INFORMATION FUSION APPROACHES

For some time, information and data fusion is established as an independent field of research. Some form of fusion is utilized in many diverse areas like robotics, pattern recognition, evidence combination as well as information retrieval (e.g. page rank aggregation). But for now there is no consistent theoretical framework developed [4].

Still, all successfully applied fusion approaches in multimedia retrieval and classification (as well as in other fields) have one common ground: they are based on relationships between the integrated modalities that can be described in form of dependency, correlation or mutual information. In general one can distinguish four basic types of information fusion: simple fusion based on data concatenation and scores, statistic and probability-based approaches and optimization with the help of information theory. Each of them will be shortly presented now.

2.1 Data concatenation and score-based fusion

The simplest and most often utilized fusion approaches are based on data concatenation and score functions applied to a vector space representation, hence to similarities or dissimilarities. Data concatenation represents

fusion on data level, where classification or retrieval indexes are calculated on the concatenated feature sets of the modalities. The classification of the full data utilizes the overall similarity between the object's features during fusion. It is a weak way of fusion since the feature concatenation is not exploiting the inter-modal relationships. Another problem is often the computational complexity due to the high dimensionality of the concatenated feature sets.

Fusion with score function is done on a more abstract level. First each modality is processed individually and then their results are fused. This can be done for classification with rule-based, ensemble methods like voting, averaging, bagging, boosting or with learning-based pattern classification algorithms like support vector machines and k-nearest neighbours. In retrieval score functions are used for rank aggregation. That is why they have a higher impact, because the fusion of different information sources (meta search) was extensively studied during the last years. The state-of-the-art approaches of diverse voting methods based on majority or position (e.g. Borda count) and multi stage approaches using the Condorect criteria can also be applied to multimedia information retrieval.

In hierarchical fusion approaches, a dimensionality reduction is achieved by processing first each modality individually. However, it maintains the modality's relationships to a certain degree in the classifier outcomes and ranked lists, which is mostly sufficient to achieve performance improvements due to fusion even at this late level. These fusion approaches exploit indirect the modality's relationships in taking their decision based on the number of modalities that support a class or the relevance of a document. Nevertheless the most of those approaches treat the modalities as independent and that is why they are suboptimal. Recently it was shown that dependent modalities are best fused by exploiting their relationships [5].

2.2 Statistic-based approaches

The approaches based on feature statistics utilize a co-occurrence matrix that is projecting the input modalities into a common semantic space. The simplest algorithm of this category is Latent Semantic Analysis (LSA) [2], more sophisticated approaches like Probabilistic LSA [10], canonical correlation analysis (CCA) [11], which uses the correlation matrix between the modalities as base, and principal component analysis (PCA) have been successfully applied as well. The co-occurrence matrix represents in the simplest case frequencies of the joint appearance of, for example, text and image features. Furthermore inverse frequencies similar to TF-IDF can be used.

The multi modal co-occurrence matrix processed with Single Value Decomposition (SVD) can then be utilized for classification and retrieval. The largest eigenvalues that are found during the decomposition represent a translation of the input modalities to a concept space, which allows clustering. When performing retrieval the query has to be translated as well to the concept space and can then be compared to the collection objects.

In those approaches the co-occurrence or correlation of features from different modalities are analysed and hence directly exploited. They are therefore efficient approaches for fusing dependent modalities.

2.3 Probabilistic-based approaches

Information fusion based on probabilistic approaches can be divided in two general variants: (1) generative modelling of modality relationships x_i and their influence to the result k (class or rank position), hence estimating the joint probability $p(x_i,k)$ and (2) discriminative modeling of the conditional probability of the result having the relationships of the modalities $p(k|x_i)$, which is derived from decision theory and circumvents the sometimes problematic task of calculating the joint probability.

Typical algorithms for this type of information fusion are mixture models [12], Bayesian networks [13] and factor graphs [14]. An extension to exploit causality (since a probabilistic relation can not be seen directly as a cause) was done with the development of causal influence networks [15]. Probability distributions can be also fused like score function using averaging and voting, which is called associative probability maps.

The probabilistic algorithms also exploit the modality relationships directly by modelling and estimating their inter-modal influences. An advantage compared to the statistic-based approaches is the possibility of modelling also uncertainty for example about the relationship strength between modalities or the amount of noise included in the modality. Furthermore, the algorithms can infer over missing data. Like the statistic-based methods they are efficient in fusing dependent modalities, because they are exploiting directly their relationships.

Another way of using probability in information fusion is estimating the accuracy of the modality's data or classification and retrieval result to down weight the influence of the weak performing ones as it is done in a lot of score-based fusion algorithms. This is done with the help of the Dempster-Shafer theory of evidence, which also provides a straightforward approach for estimating the relevance of a document to a query.

2.4 information theoretic optimization

Information theory is not directly utilized for fusion, but for optimizing other information fusion algorithms as the ones mentioned above. The goal is to minimize the uncertainty about the fusion result during fusion, which can be measured with the conditional Shannon entropy of the result having the inputs $H(k|x_i)$. At the same time this maximizes the mutual information between input and result $I(x_i,k)$. This entropy fusion model is then used for feature selection in the information fusion approach of choice [8]. In this way an optimal feature subset can be determined that contains the most information about the treated problem. Therefore information theory is an important tool in information fusion.

As this review shows all information fusion approaches exploit the dependency, co-occurrence, correlation or mutual information between modality features. The next section will discuss relationships in terms of how to determine them and which amounts can be found in data collections like Corel and Washington. Furthermore our findings are used to predict problems in the application of current state-of-the-art fusion algorithms to web content as can be found in the Wikipedia collection.

3. THE BASE OF INFORMATION FUSION: RELATIONSHIPS BETWEEN MODALITIES

All types of information fusion are based on exploiting some form of relationship between the integrated modalities like dependencies, co-occurrence, correlation or mutual information. So for multimedia information retrieval the basic idea is to use the relationship between the textual and visual feature sets to cluster the keyword-annotated images or perform multi modal based retrieval, which should also make it possible to retrieve un-annotated images with text queries that are visually similar to annotated images that match the query.

Another application area is the classification and retrieval of websites merging all available types of media source like texts, images, videos and audio files, their transcripts and even the meta data, the website's structure and links. According to information theory the exploitation of more information is leading to a performance improvement in retrieval and classification, if the fusion of the dependent modalities can be performed appropriately. But this task is not trivial, since incorrect merging of dependent inputs will hurt the system's overall performance.

That means in the practice of information fusion that it is important to study the relationships that are underlying the examined data. For our first preliminary fusion tests based on visual and textual feature the two keyword annotated image collections Washington and Corel were chosen. The Washington collection contains 675 images clustered in 16 semantic concepts (classes with missing annotation were discarded) and having each 1-5 keywords. The images were annotated manually and the images are equally distributed over the concepts. The second collection, a subset of the Corel database, has more images (1159) with more keywords (1-10) and are clustered into 49 concepts. Contrary to Washington the images are not evenly distributed over the classes and the annotations also contain complete non-sense descriptions. So it can be considered to resemble more to real world data as one has to expect for example in web content. The analysis of the Wikipedia collection to prove this prediction was not done yet. The assignment of the images to their semantic concepts was used as groundtruth in the following experiments.

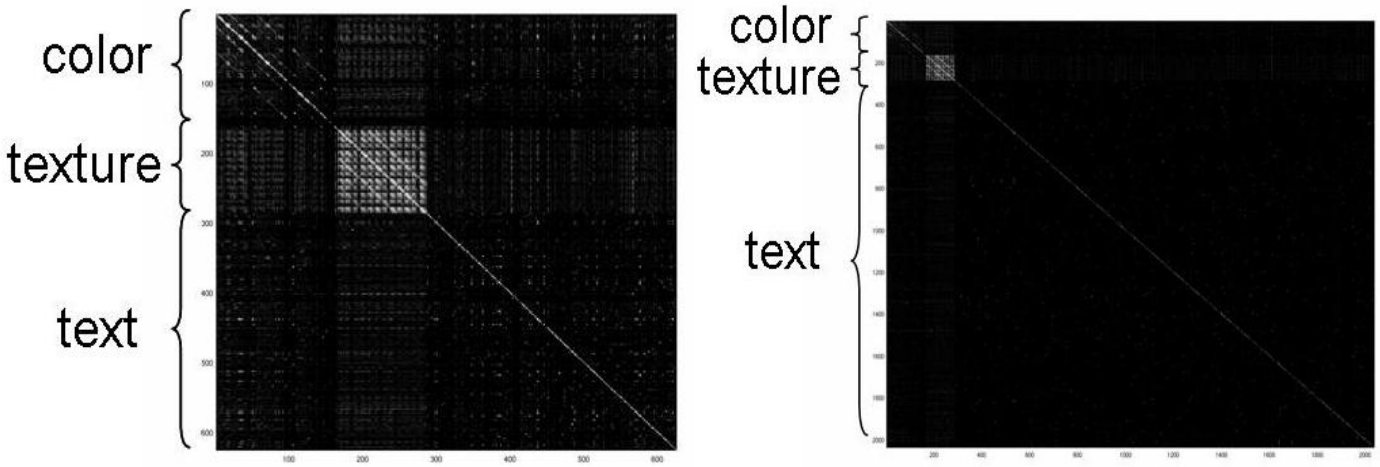

FIGURE 1: Absolute Correlation coefficient matrices between the features of (left) Washington collection and (right) Corel collection

In preparation of running the correlation and fusion experiments for both collections textual (feature vector of term frequencies Washington: 338, Corel: 2035, very sparse) and visual features (Gift features [6] color/texure histograms: 166/120) were calculated. In Figure 1 the absolute correlation coefficient matrices between the features of both collections are presented. Light points mark high correlation (positive or negative) and dark points represent small correlation and hence independence of the features.

It is clearly visible that with the rise in size and resemblance to real world data from the Washington to the Corel collection a smaller amount of correlation is obtained between their features. More precisely, the Washington collection contained 17% significantly correlated feature pairs, whereas there were only found 3% in the Corel collection. Here, it has to be taken into account that the texture features are highly correlated by their nature (wavelet coefficients), which reduces the amount of useful dependencies further.

Table 1 shows additionally the absolute average correlation coefficient and in brackets its maximum that has been found between the modalities in the Washington and Corel collection respectively. It can be seen that the inter modal dependencies are propagated according to the maximum to the classifier outcomes, which have been calculated with a standard support vector machine (SVM) such as later used for the fusion experiments. The impression that the color-texture relationship inheres the most dependence (and hence is most valuable in fusion) is skewed by the difference in sparsity of the data.

C	*Washington*			*Corel*		
	color-texture	***color-text***	***texture-text***	***color-texture***	***color-text***	***texture-text***
average	0.077 (0.542)	0.035 (0.857)	0.049 (0.738)	0.043 (0.434)	0.015 (0.991)	0.025 (0.804)
classifier	0.275	0.516	0.268	0.141	0.211	0.199

TABLE 1: Average, maximum and minimum correlation coefficients between features found in Washington and Corel collection; correlation coefficient found between the classifier outcomes trained on modalities

The small amount of feature dependencies in the Corel collection leads to the assumption that the Wikipedia collection, where the images are aligned with free texts, inheres more noise and probably even less dependencies than the investigated Corel subset. The significantly correlated feature pairs in the overall collection will probably be close to zero. This finding coincides with reports, whereas the content of texts in websites is not always a helpful description for the aligned images. This has a great impact on the choice of approach for solving this retrieval and classification problem, since all algorithms that were presented in section 2 are based feature dependencies.

Maybe it is possible to alleviate the problem of small dependencies by using more sophisticated features for both images and text. Furthermore, the consideration of intra-class dependencies (which features discriminate a certain query/class best against the rest of the collection) seems promising in such huge and noisy collections.

in%	***color***	***texture***	***text***	***hierarchical SVM***	***concatenated SVM***	***averaging***	***weighted sum***	***majority vote***
classification error	34.4	37.4	26.5	**3.7**	46.7	31.6	29.6	27.9
false alarm	6.9	25.4	**1.1**	40.6	14.1	2.1	1.7	4.1
Miss	35.5	37.9	27.5	**2.1**	47.7	32.8	30.7	28.9

TABLE 2: Information fusion performance of the Washington collection

Fusion experiments on the Washington and Corel collection show that the difference in the amount of dependencies has a direct impact on the information fusion performance as can be seen in Tables 2 and 3. Therein the classification error, the false alarm rate (false negative) and miss rate (false positive) of the single modality classification (color, texture, text) with a standard support vector machine (SVM) and several information fusion approaches is given. All the tested approaches belong to the data concatenation (concatenated SVM) or score-based fusion algorithms. We investigated rule-based ones (averaging, weighted sum, majority vote) and a learning-based algorithm (hierarchical SVM). More experiments concerning the impact of fusing dependent and independent features can be found in [3].

in%	***color***	***texture***	***text***	***hierarchical SVM***	***concatenated SVM***	***averaging***	***weighted sum***	***majority vote***
classification error	45.3	47.7	45.5	**15.4**	46.9	47.2	46.8	44.8
false alarm	26.5	37.9	20.3	49.5	23.8	44.8	**18.5**	21.8
Miss	45.6	47.9	45.9	**14.7**	47.3	47.6	47.3	45.2

TABLE 3: Information fusion performance of the Corel collection

The above experiments show that the fusion of the less accurate and as well less correlated classifier outcomes (since they are obtained from less correlated data) results in a decreased performance in all types of information fusion. Which one of the two factors is influencing more this decrease can not be stated from these experiments, but would be also important to determine in a future task. The worse results of the Corel collection further underpin the problem that researchers are faced with when fusing even less correlated data, as expected for the real world data of the Wikipedia collection.

4. CONCLUSIONS AND FUTURE WORK

This article presents a possible explanation why multimedia retrieval and classification with huge real world data collections like web content stays for now behind the expectations that, in theory, the fusion of more information should lead intuitively to improved performance. If this data contains too little dependencies between the modality features or most of the dependencies are hidden in noise, then all standard information fusion approaches are preassigned to fail, since they are based on those relations.

Our future work will first concentrate on the dependency analysis of the Wikipedia collection. The calculation of the overall and intra class correlation coefficient matrices will show if the predictions made earlier for the Wikipedia collection hold true. Here the intra-class dependencies will be especially interesting, since each class possess most certainly their own set of features that distinguish them best from other classes. Those dependencies are probably then also less sensitive to the collection size and its noise level, since it is based on a small subset of the data.

In addition to the presented correlation coefficients, other dependency and overlap measures will be examined for their expressiveness of the feature relationships. For example, a promising one would be the pattern magnity coefficient that takes, except of the features co-variances, also their magnitudes into account. More appropriate measures are also the Jaccard index (represents similarity and diversity of sample sets), the Dice coefficient, Saulton's cosine measure or the data overlap coefficient, which can be applied to discrete data. All those measures are derived from information theory. Alternatively one can also examine the mutual information of the different modalities.

Another research direction will be the exploitation of more sophisticated features for the text as well as the images. For text for example inverse frequencies seem promising. In general, better natural language processing like finding named entities can be very helpful. Finally, the structure of the websites itself as well as the link structure can be exploited. The idea is that in text that is closer to the image better describing keywords can be found.

The currently utilized image features also leave a lot of room for improvement, since for now only the most simple, global color and texture histograms are applied. A first step of improvement would be to include Gift's local color and texture features. We also like to test the performance of the MPEG7 [9] features in terms of achieving better dependencies with the textual features.

The logical next step would be the study of region-based indexing approaches. Intuitively, they raise the chances of being related to keywords found in the image's aligned texts. The final step for the image processing would be the utilization of object recognition or as a preliminary step detecting semantic concepts in the images such as faces, buildings, nature and so on.

Once appropriate features have been found, which includes a measure for the inter-modal dependencies, the strength of their relations can be determined. Depending on this amount one of the standard fusion approaches presented in section 2 can be applied or another solution has to be found to handle extremely noisy data.

ACKNOWLEDGEMENTS

This work was supported and financed by the European project MULTIMATCH. I also like to thank my supervisors Eric Bruno and Stephane Marchand-Maillet for their support and fruitful discussions.

REFERENCES

[1] van Zwol, R; Kazai, G. and Lalmas, M.. (2005) INEX Multimedia Track.

[2] Deerwester, S.; Dumais, T. D. and Harshman, R.. (1990) Indexing by Semantic Analysis. Journal of the American Society of Information Science , 41 , 391-407.

[3] Kludas, J.; Bruno, E. and Marchand-Maillet, S. (2007) Information Fusion in Multimedia Information Retrieval. 5th International Workshop on Adaptive Multimedia Retrieval (AMR), July 2007 (to appear).

[4] Kokar, M. M.; Weyman, J. and Tomasik, J. A. (2004) Formalizing classes of information fusion systems. Information Fusion, 5, 189-202.

[5] Koval, O.; Voloshynovskiy, S. and Pun, T. (2007) Error exponent analysis of person identification based on fusion of dependent/independent modalities. Proceedings of SPIE-IS&T Electronic Imaging 2007, Security, Steganography, and Watermarking of Multimedia Contents IX.

[6] Squire, D. M.; Müller, W; Müller, H. and Raki, J. (1999) Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. (143-149)

[7] Smeulders, A. W. M.; Worring, M.; Santini S.; Gupta, A. and Jain R. (2000) Content-Based Image Retrieval at the End of the Early Years. IEEE Transaction on Pattern Analysis, 22(12), 1349-80.

[8] Fassinut-Mombat, B. and Choquel, J.-B. (2004) A new probabilistic and entropy fusion approach for management of information sources. Information Fusion, 5, 34-47.

[9] Manjunath, B.S.; Salembier, P. and Sikora, T. (2002) Introduction to MPEG-7: Multimedia Content Description Interface. Wiley & Sons.

[10] Hofmann, T.; (1999) Probabilistic latent semantic analysis. In Proceedings of the 15th Conference on Uncertainty in AI, 289–296.

[11] Fortuna, B. (2004) Kernel canonical correlation analysis with applications, SIKDD 2004 at Multiconference IS.

[12] Westerveld, T.; de Vries, A. P. (2004) Multimedia Retrieval using Multiple Examples, Proceedings of the International Conference on Image and Video Retrieval (CIVR), 344-352.

[13] Jasinschi, R.S.; Dimitrova, N.; McGee, T.; Agnihotri, L.; Zimmerman, J.; Li, D. (2001) Integrated multimedia processing for topic segmentation andclassification, Image Processing, 3, 366 – 369.

[14] Ramesh Naphade, M.; Kozintsev, I.V.; Huang, T.S. (2002) Factor graph framework for semantic video indexing, Circuits and Systems for Video Technology, 12, 40 – 52.

[15] Wu, Y.; Chang E. Y. and Tseng B. T. (2005) MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia, 872—881.