# A Multimodal Fusion Approach By Exploiting Concept Interactions for Efficient Multimedia Analysis

Elvan Gulen
Department of Computer Engineering
Middle East Technical University
06800 Ankara, Turkey
gulen@ceng.metu.edu.tr

Turgay Yilmaz[*]
Department of Computer Engineering
Middle East Technical University
06800 Ankara, Turkey
turgay@ceng.metu.edu.tr

Adnan Yazici
Department of Computer Engineering
Middle East Technical University
06800 Ankara, Turkey
yazici@ceng.metu.edu.tr

## ABSTRACT

Multimedia data intrinsically contains multimodal information in it. In order to obtain a successful multimedia analysis, all available information should be utilized by following a multimodal approach. In addition, the interaction between concepts is another crucial source of information and helps to increase the fusion performance. The focus of this study is to show that fusing all available modalities along with the concept interactions can yield better results. The experiments conducted on TRECVID 2007 dataset validated the superiority of such combination over best single modality and alternative modality combinations. Combining all available modalities performs 10.5% relatively better than the best single modality in overall. Additionally, utilizing from other concept information provides a 5.6% relative performance gain over the multimodal fusion results.

## Keywords

multimedia content analysis, multimodal fusion, semantic concept detection, concept interactions

## 1. INTRODUCTION

Increase in the use of digital video data in recent years has exposed the need for content based video retrieval systems. Content based retrieval of videos requires extracting the semantic content in videos. However, the 'semantic gap' between the low-level features of video data and the high-level semantic information is still a challenging problem. Thus, semantic content extraction is still an attractive topic for the researchers [8].

In order to provide an effective semantic content extraction solution, the nature of the multimedia data should be analyzed carefully and the contained information should be used throughly. Video data exhibits an unstructured characteristic and causes several complexities like lighting variations, viewpoints changes, camera motion, occlusion, noise in the sensed data, etc. Besides, video data has another significant characteristic, which can help to overcome these challenges; the multimodal content. Combining the information gathered from multiple modalities is an empirically validated approach to increase the retrieval accuracy [2]. Moreover, the dependence on any modality can be minimized through fusion and this yields to a more robust system. We can think of a *People-Marching* event as an example, where the event can be recognized by using in any of the visual, auditory and textual modalities. The video can include people as visual objects, a shouting sound and also some lyrics of a march in the closed caption text. A combination of these modalities can provide a higher detection accuracy and is less dependent on potential problems in any of the modalities.

The information fusion literature contains a significant number of studies on multimodal fusion. However, most of these studies do not benefit from all available modalities and focus on some alternative modality couples, especially 'audio-visual' and 'visual-textual' [2, 14]. In this study, we would like to incorporate as much information as possible via the available modalities. Thus, we accept that a 'modality' is a set of information which is complementary to the other included modalities [29] and elaborate the three information channels in video (visual, auditory, textual) into following complementary modalities: Visual-Color, Visual-Region, Visual-Texture, Audio-Perceptual, Audio-Cepstral and Textual. Thus, we try to make use of any useful information included in the video data and increase the retrieval accuracy.

Different from the modalities combined, another crucial source of information is the interaction between the concepts. Considering that videos can contain more than one concept (i.e. objects, events) at a time, the co-occurrence information obtained via other concepts can assist as an

[*] Turgay Yilmaz is currently with the Center for Information Fusion, Institute of Industrial Science, University of Tokyo, Japan.

additional cue to increase the detection performance. For instance, it is obvious that a *Boat_Ship* concept usually appears with *Sea* and *Harbors* concepts. So the interaction information between *Boat_Ship* and these concepts can be helpful for detecting the *Boat_Ship* concept. However, such information is usually ignored by many of the fusion approaches [13, 18]. Considering a late fusion scheme, which combines the classifier output scores, only the corresponding output score of each classifier with the retrieval class is used during combination. For instance, the fusion result of *Boat_Ship* is calculated by using only the *Boat_Ship* concept scores of each modality. In this study, in order to benefit from the information that the interaction between concepts provides, we utilize all score outputs of all modalities while combining them.

In addition to the issues depicted, a crucial evaluation on the currently available multimodal fusion approaches is that they are usually domain specific [2]. The combination rules and approaches are strictly bound to particular domains, which causes scalability problems. In this study, we do not limit the system to any particular domain and propose a generic multimodal combination approach.

Considering the above given aspects, the contribution of this study is three-fold. First, we try to enhance the accuracy by employing as much complementary modalities as possible. Second, we utilize the interactions between concepts. Third, we implement a generic multimodal combination approach. We conduct several experiments on the TRECVID-2007 benchmark, and compare the combination results with single modality accuracies and alternative combinations of modalities, in detail. We conclude that the proposed concept-interaction based multimodal combination approach performs superior than using single modalities or any couples of modalities.

The remainder of this paper is organized as follows: In Section 2, the analysis of the literature on multimodal information fusion is given from various aspects. In Section 3, the proposed multimodal and concept interaction based fusion approach is given in detail. In Section 4, the incorporated modalities are presented and the systematical approach for integrating them is given. Besides, the empirical results and the evaluations are presented. Lastly, in Section 5, some conclusions are drawn and further study is discussed.

## 2. RELATED WORK

In the multimedia analysis literature, a big majority of the studies focus on a single modality such as images, audio or text [16, 25, 31]. Besides, a significant number of studies in the multimodal information fusion literature use two modalities; visual and auditory, or visual and textual, etc. In [32], Zhu et al. combine visual and textual information for image categorization and show that adding textual information definitely improves the results. However, using all available information provides more valuable information when the combined modalities include complementary information. For instance; in [9], combining visual and auditory evidences increases the best unimodal performance; when the textual modality is added to the fusion process, 13% performance improvement is attained. In addition, according to the experiments conducted in [1], the highest retrieval performance is reached when all modalities are combined. Yet, the definition of modality is still vague and it can be concertized in different ways. At one extreme, each of the features can be treated as separate modalities. At the other extreme, all of the features can be treated as one modality [29]. A mid-point can be grouping the features according to some criteria. For instance; it can be regarded that the extracted features of a video object can be grouped according to their media-source: formulating visual, auditory, and caption modalities with related features. However, each of these modalities can be expanded. For instance; even visual data can be defined with several modalities like color, shape, texture, face, etc. In this study, we identify six modalities for video data (Visual-Color, Visual-Region, Visual-Texture, Audio-Perceptual, Audio-Cepstral and Textual), by considering the different information they contain.

Another crucial issue in the design of a multimodal fusion system is the scalability issue. Most of the studies in the field of information fusion are domain specific. The predominant research domains are sports and news. The studies usually focus on very specific tasks like finding certain concepts or segmenting videos into stories and predefined genres. In [12], fusion is utilized to categorize new videos into health, sports, finance, politics and society. In [20], visual and auditory cues are incorporated to reach the goal event in sports videos. Additionally, some of the domain specific studies prefer knowledge based fusion approaches since knowledge based rule systems usually perform well for simple and specific tasks [4, 20, 26]. As an exmaple, in [26], knowledge based rules use visual and auditory cues to identify the shots with/without speaker, shots with/without person, and to detect video parts including silence/speech. However such methods are lacking in terms of scalability and robustness. In order to deal with these shortcomings, researches lean to machine learning approaches [23] such as Bayesian Networks, as in [6, 30], and SVM, as in [3, 22, 24], etc.

Yet, studies having the ability to process different contents from various domains are low in number. However, it is crucial to yield a wider coverage for more effective retrieval of multimedia data. If the studies which are relatively generic are considered, it can be seen that they do not benefit from the interaction between concepts, though their reasonable performance in accuracy [24, 28]. However, since concepts may have positive or negative associations, this knowledge may play an useful role in increasing their concept detection accuracy. In this study, we propose a generic multimodal combination approach that tries to utilize as much modalities as possible and benefits from the interaction of concepts.

## 3. MULTIMODAL INFORMATION FUSION

The proposed fusion system is designed to work in cooperation with other single modality-based systems and enables performing the semantic content analysis task in a multimodal manner. Consequently, the system enables producing semantic information ready to be stored in a multimedia database for further retrieval tasks. In our fusion method, a score-based late fusion scheme is preferred considering its favorable performance compared to the early fusion approaches [24]. Additionally, the score-based late fusion approach is more appropriate for modeling concept relations among different information sources. We aim to build the combination system as generic as possible and domain independent so that it can be expandable with new concept definitions.

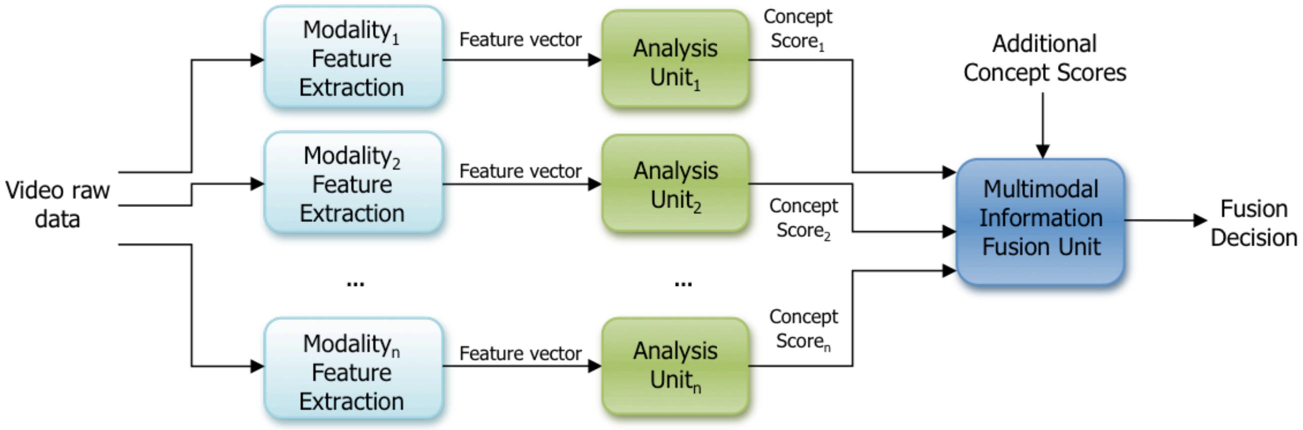

**Figure 1: The general scheme of the fusion approach**

General architecture is presented in Figure 1. As illustrated, all the classifiers, i.e. analysis units, work with a single feature type. Each of these single modality based classifiers perform binary classification to output the detection score of the related concept. Assuming that we have $n$ number of modalities, and $m$ number of concepts in the concept lexicon, the process flow in Figure 1 is repeated for all $m$ concepts.

After obtaining detection scores from each classifier, these scores are preprocessed in order to ensure the synchronization between different modalities. The temporal alignment between all available modalities is performed with respect to the visual shot boundaries, regarding that most of the concepts are highly connected to the visual content of the video. For each shot, the detection scores of the target concept and other concepts appear in the related shot constitute the input of the fusion system. Support Vector Machine (SVM), one of the most successful classification approaches [5], is chosen as the fusion method. It is preferred due to its improved classification performance, and ability to handle unbalanced or noisy data. In addition, it is observed to be more effective and mostly used in studies which follow a multimodal approach for the semantic concept detection task [1, 3, 10, 28]. For SVM implementation, libSVM [7] is used.

After modality-based data processing, the fusion phase takes place and the resultant scores gathered from several modalities are combined into a multimodal representation to yield a final detection score for each concept. In our research, we extend such multimodal fusion approach by employing additional semantic cues (i.e. detection scores of other concepts). To set the distinction between these two, we refer the first fusion scheme as multimodal fusion and the second one as interaction-based multimodal fusion rest of the paper.

## 3.1 Exploiting Concept Interactions

As mentioned before, interactions between different concepts are usually not considered by many fusion strategies. For instance; most popular fusion operations like simple averaging, minimum/maximum selection and weighted linear combination do not consider such an interaction. They combine scores or decisions of the same target class. However, since we want to exploit the interactions between concepts, our fusion system is designed to incorporate the scores of other concepts into the fusion process, which can be accepted as additional semantic cues for information fusion. As the additional semantic cues, it may be possible for the fusion system to use the outputs of own classifiers for other concepts or take them from an external source (as performed in the experiments).

In interaction-based multimodal fusion, to make better prediction of concepts, we enhance the fusion process by utilizing other semantic cues. Some concepts usually cooccur together and some other are most likely not present at the same time. Considering the association between concepts, we propose that additional semantic information can be helpful to make a better prediction of concepts. Thus, we expand our fusion approach by using the semantic cues belonging to other concepts, in addition to combining defined six modalities. However, concept detection task is usually evaluated through a small set of concepts because of the lack of resources and annotated data. In order to obtain more successful results a larger set of concept scores could be more helpful. Considering the benchmark we use in our experiments, for each target concept we could be using the scores of the remaining 19 concepts. However, these 20 concepts are mostly unrelated, so they can be weak to provide valuable semantic information. Besides, they don't appear together in most of the shots, so we need a large concept lexicon including more concept cues with more positive relations along with negative ones. For this task, we adopted two popular benchmarks, VIREO-374 and Columbia-374. We use the average fusion results of their detection scores on the lexicon of 374 semantic LSCOM concepts as additional semantic cues. So for each concept, 373 of the detection scores (i.e. concept scores of other than the target concept) of CUVIREO-374 [11] are used. The key point behind choosing this lexicon is that it contains a large number of concepts and also includes concepts related to each TRECVID 2007 concepts. For instance; the *Smoke* concept in the LSCOM lexicon, can be helpful to detect the *Explosion_Fire* concept more accurately, as well as *People-Marching* can benefit from the *Crowd* and *Protesters* concepts.

## 3.2 Fusion Strategy

For each concept, the fusion approach constructs a SVM model which is later used for estimating the probability whether a new shot contains the target concept or not. For building SVM models, ground truth of each shot in the training set is used and the feature vector for each shot is linked with these labels. The concept classifiers are trained with the corresponding feature vectors which are composed of scores obtained from single modalities, e.g. predictions of the target concept from different experts and detection scores of other concepts.

The training procedure includes several steps. First, considering different modalities can have different score intervals, the training data is scaled into [0,1] interval and then the model parameters are determined. After a few evaluations between popular kernel methods (sigmoid, linear, RBF, etc.), Radial Basis Function (RBF) kernel is chosen because it is found to be more effective to model the concepts. Since the classification performance of SVM can highly be dependent on the model parameters, finding the optimal parameter combination for each concept classifier is crucial. Since RBF is used, we need to identify a good combination of the RBF parameter $\gamma$ and the soft margin parameter $C$ so that the classifier can successfully predict the presence of the target concept in the test shot. So we perform a grid search in the parameter space with k-fold cross validation to optimize these parameters. In cross validation procedure, the training data is randomly divided into k equal subsets and each subset is tested by the classifier trained on the remaining subsets. Since cross-validation time complexity increases according to the value of k, we choose k as 10 which is successful to find a good parameter combination within a reasonable time. Grid search conducts a search through a subset of the parameter space with exponentially growing sequences to find the best parameter values. So each combination of parameter choices is checked using cross-validation, and the parameters maximizes the performance are picked. At this point, to evaluate the performance, the procedure must be guided by a performance metric. Mostly, accuracy metric is used for this purpose and also libSVM supports just the accuracy metric. However, it may not work very well in certain cases. Considering the unbalanced datasets, accuracy may not yield good parameter combinations because even it predicts all samples as the dominant class, the accuracy can still be very high. To prevent this problem one solution can be balancing the training data. However, when the number of positive samples are very low, and if same number of negative samples is taken as positive samples, it won't be very successful. For instance; there are concepts having positive samples below 100 shots in thousands of shots in the benchmark used for evaluations. Moreover, positive samples of a specific concept among all video shots are usually very low in real life cases. Also we want to use the whole available information. So in order to avoid the imbalanced ratio problem, we developed an extension to libSVM which enables to perform cross-validation under different criteria such as f-measure, accuracy, etc. In the learning phase of fusion process, since the minority class is more important, we perform cross validation with f-measure criteria. After model selection, SVM is trained with these parameters and a SVM model specific for the target concept is constructed.

In testing phase, when an unlabeled shot is asked, each concept classifier is tested with the appropriate feature vector. Then it provides a detection score according to the presence of the concept in the shot. SVM is normally a binary classification method but since we need scores to evaluate the retrieval performance, the output of SVM is converted to probabilistic outputs by Platt's method [21].

# 4. EMPIRICAL STUDY

In this section, we evaluate the performance of the porposed multimodal and interaction-based multimodal fusion schemes. The experiments are conducted on the TRECVID 2007 [19] dataset which includes 100 hours of videos from various domains such as documentaries, news reports, science news, etc. We build concept detectors for the 20 officially selected concepts for the TRECVID 2007 evaluation. There are 21532 reference shots for the training set and 18142 shots for the test data in total. Since the training phase took so long when using all available training shots, we choose a subset of them. For all concepts, the number of the negative samples are a lot more than the positive samples in the training data. So while building up the training subset for each concept detector, we choose all available positive samples and randomly select from negative samples until it contains information of 2500 shots. In evaluation, *Average Precision (AP)* metric is used to report the performance on individual concepts and *Mean Average Precision (MAP)* for the overall system performance, respectively. While measuring the retrieval accuracy, $AP$ is calculated at a depth of 2000 which is predetermined by TRECVID benchmark. Further details about the sataset and a performance comparison of TRECVID 2007 participants can be found in [19].

For semantic video analysis, the video should be segmented into subunits so that the low-level features can be extracted and processed to build semantic concept models. We consider the main expertise as the visual modality. We follow a simple synchronization at shot level where the auditory and textual features are extracted according to the time of the shot boundaries. For each shot, the middle keyframe is chosen as the representative keyframe from where the visual features are extracted. Besides, since we work on TRECVID benchmark, we stick with their provided shots which were extracted from the visual part. All information sources of a video, i.e. sound, images, texts, are processed for semantic content analysis and selected features are extracted from these sources.

All features are categorized into six modalities with respect to the information type they contain. For visual modalities, MPEG-7 features are extracted by the MPEG-7 reference software (eXperimentation Model, XM) [17] and are grouped into three modalities which are color-based modality (Color Layout, Color Structure, Dominant Color and Scalable Color features), shape-based modality (Contour and Region Shape features), and texture-based modality (Edge Histogram and Homogeneous Texture features). We have formed two auditory modalities; the first modality includes Energy, Linear Predictor Coefficients, and Zero Crossing Rate features referred as spectral-based, and the other modality includes Mel-frequency Cepstrum Coefficients. These auditory features are extracted with Yaafe toolbox [15]. Finally, for the textual modality, the term frequency-inverse document frequency (TF-IDF) is calculated from the Automatic Speech Recognition (ASR) and Machine Translation (MT) texts. Since the feature dimension is very large for the textual modality, a dimensionality reduction technique,

**Table 1: Evaluation results of single and combined modalities for detecting TRECVID 2007 concepts. Best performance for each concept is given in bold.**

| | Visual Color | Visual Texture | Visual Shape | Audio Perceptual | Audio Cepstral | Textual | Multimodal | Interaction Based Multimodal |
|---|---|---|---|---|---|---|---|---|
| Airplane | 6.55% | 8.25% | 6.35% | 7.33% | 5.11% | 6.09% | **9.83**% | 9.36% |
| Animal | 14.34% | 17.18% | 5.96% | 8.51% | 9.36% | 7.76% | 17.09% | **17.65**% |
| Boat_Ship | 6.40% | 12.79% | 8.97% | 7.26% | 8.77% | 8.14% | 14.86% | **16.80**% |
| Car | 15.63% | 19.75% | 14.25% | 14.43% | 16.18% | 10.42% | 21.38% | **23.30**% |
| Charts | 3.37% | **5.95**% | 2.42% | 2.51% | 1.34% | 1.87% | 4.52% | 4.98% |
| Computer_TV-screen | **12.74**% | 9.60% | 7.42% | 5.38% | 8.13% | 8.23% | 10.83% | 11.15% |
| Desert | 1.92% | 2.00% | **2.86**% | 0.78% | 1.04% | 0.57% | 2.03% | 2.26% |
| Explosion_Fire | 1.96% | 2.32% | 1.75% | 2.47% | 1.40% | 2.11% | 1.66% | **2.52**% |
| Flag-US | 0.10% | 0.42% | 0.26% | 0.13% | **1.61**% | 0.25% | 0.17% | 0.31% |
| Maps | 6.36% | 10.94% | 4.10% | 2.61% | 4.68% | 2.47% | 10.72% | **11.52**% |
| Meeting | 23.69% | 24.55% | 23.75% | 23.76% | 29.41% | 19.12% | 31.16% | **31.44**% |
| Military | **5.13**% | 3.14% | 1.94% | 3.45% | 0.96% | 0.99% | 3.69% | 3.87% |
| Mountain | 8.92% | 5.90% | 3.86% | **9.06**% | 2.70% | 2.74% | 6.46% | 7.59% |
| Office | 8.99% | 13.90% | 7.09% | 5.63% | 9.49% | 2.66% | 13.42% | **14.71**% |
| People-Marching | 7.00% | 7.42% | 2.97% | 3.37% | 3.80% | 1.45% | 8.34% | **9.86**% |
| Police_Security | 1.64% | 4.87% | 2.98% | 3.14% | 2.89% | 2.52% | **4.96**% | 4.94% |
| Sports | **11.08**% | 5.50% | 3.36% | 3.67% | 4.58% | 1.75% | 7.58% | 9.52% |
| Truck | 7.97% | **10.90**% | 5.47% | 5.07% | 7.51% | 5.77% | 10.61% | 9.24% |
| Waterscape_Waterfront | 16.75% | 17.54% | 16.31% | 9.44% | 10.57% | 8.21% | **22.48**% | 22.22% |
| Weather | 1.62% | 1.38% | 0.35% | 1.80% | 0.30% | 0.27% | 1.84% | **1.87**% |
| MAP | 8.11% | 9.21% | 6.12% | 5.99% | 6.49% | 4.67% | 10.18% | **10.76**% |
| MAP Rank | 4 | 3 | 6 | 7 | 5 | 8 | 2 | 1 |
| # Of Best Ranks | 3 | 2 | 1 | 1 | 1 | 0 | 3 | 9 |
| Average Rank | 4.45 | 3.1 | 5.8 | 5.5 | 5.65 | 7 | 2.8 | 1.7 |

i.e. Diffusion Maps, is adopted to reduce the feature space into lower dimensions. A matlab toolbox for this purpose, i.e. dimensionality reduction, is utilized [27].

For each modality, each shot is represented as a feature vector. These feature vectors are fed into the learner to model the target concept. SVM is chosen for the concept modeling because of the previously mentioned reasons in Section 3. Therefore, a SVM model is constructed for each concept. In this part, six SVMs are learned separately for each concept and tested to construct individual detection scores to be further given as an input to the fusion process.

In Table 1, the retrieval performances of unimodal approaches and proposed fusion schemes are compared. The results indicate that fusing multiple information outperforms any single modality based approaches in overall. Additionally, we see that an important performance gain is obtained by fusion for most of the concepts (12 of 20 concepts). Also it gives comparable results for the remaining concepts. The proposed multimodal fusion approach provides 10.5% and interaction-based multimodal fusion approach provides 16.7% improvement over the best unimodal baseline. Since interaction based multimodal fusion results in an improvement of 5.6% over multimodal fusion, we can conclude that additional semantic cues play a key role in contributing to increase the performance.

Most of the concepts benefit from multimodal fusion with additional concept cues. For example; *Boat_Ship* detection performance increases from 12.8% for the best single modality baseline to 14.9% for multimodal fusion. In other words, multimodal fusion provides a 16.2% performance gain in detecting *Boat_Ship*. For some concepts a further improvement is obtained by feeding the information of other concepts, i.e. Cu-vireo374 detection scores, into the fusion process. For instance; retrieval performance of *People-Marching* enhances the performance from 8.3% with multimodal fusion to 9.9% with additional concept cues. Similarly, retrieval performance of *Car* jumps from 19.7% to 21.4% through multimodal fusion and goes up to 23.3% by interaction-based multimodal fusion. On the other hand, fusion doesn't provide an improvement of some concepts such as *Charts* where the texture based modality baseline shows the highest result. The reason why fusion doesn't perform better in detecting *Charts* concept may arise from the dominance of the textural structure of the concept. Besides, other modalities may not contain valuable information of that specific concept.

Apart from handling each six modality separately, we perform more tests to see the effects of visual, auditory modalities and modality couples, respectively. To obtain one result for each concept from the visual cues, we integrate the results of the three visual-based modalities with the same fusion method. Similarly, we fuse the detection scores of the auditory-based modalities to reach one auditory-based score. Furthermore, we apply our fusion strategy on these modalities to get the binary combinations of visual, auditory and textual modalities. In Figure 2, the average precision results of three modalities, couple modalities, fusion of all of them and interaction-based multimodal fusion are given. Note that, the detection scores obtained from the six modalities aren't shown in the figure. First of all, we observe that applying fusion on the three visual modality and the two auditory modality, increases the retrieval performance for most of the concepts along with the overall performance. Also, binary combinations provide performance gain over these modalities. Even, interaction-based multimodal fu-

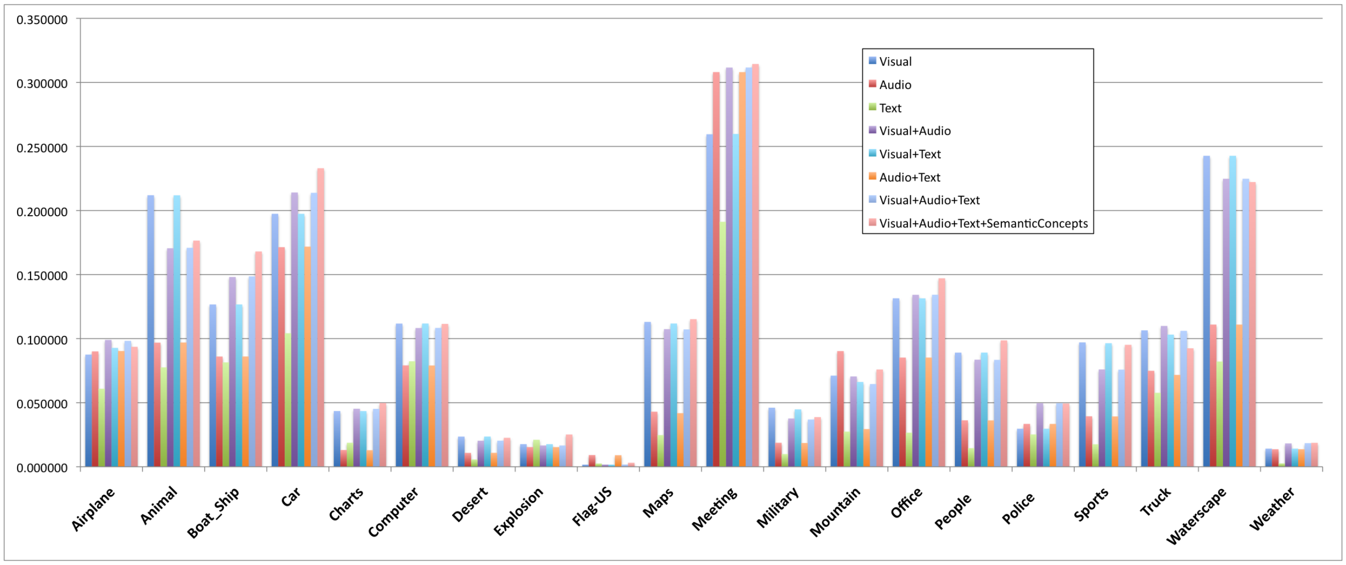

**Figure 2: Average precision comparison of various runs**

sion mostly performs better than all runs, the fusion results obtained without using additional cues doesn't show significant performance gain over the best modality couple (visual+auditory). For the greater part of the concepts fusing visual and auditory cues performs slightly better than fusing all. Therefore, these close and comparable results indicate that the textual modality doesn't make a major contribution to the fusion process. The leading cause of it is that the textual features we use in our tests aren't very accurate. As previously mentioned, the textual features are the ASR texts of the TRECVID 2007 videos. Since the majority of these videos are non-English, the ASR output later translates into English by MT and this may yield a substantial data loss. As a result the accuracy of the textual data can be low. Instead of using the translation text of the ASR output, obtaining potential textual meta data related to the videos and adding them to the fusion process can presumably yield more successful results.

The success of the fusion system is highly dependent on the quality of the input information, i.e. unimodal detection scores. So in order to obtain higher results, more successful independent concept detectors can be built. For this purpose, the performance of other popular and successful features can be investigated such as SIFT. Moreover, the low performance of the textual modality can be increased by using other textual meta data as features and considering the texts appearing in the neighboring shots. However, since the study emphasizes on the fusion part, we need to consider possible extensions accordingly. In this research, for each SVM concept model, we use RBF kernel, but other kernels may be better in learning certain concepts. So a kernel method selection procedure can be included to find the right kernel for each concept model. After a few evaluations, we observe that selecting all available information results better than selecting a small set. The results suggest that for each concept, some modalities are more important than others and some modalities may not provide valuable information about that concept. Hence, a powerful feature selection procedure to select good modalities and concepts having strong relations to the target concept can be helpful to obtain better fusion performance.

## 5. CONCLUSION

In this paper, we propose a general fusion system to obtain more accurate results for semantic content analysis task and the contributions of fusing several information sources is also investigated. The experimental results show that the proposed fusion approach outperforms other single-modality based approaches. Therefore, we conclude that performing multimedia content analysis by exploiting multimodal information shows a significant improvement in the performance of the overall system. Besides, the proposed idea of gaining benefit from concept relations, in other words the positive effect of using other concept information is verified.

## 6. ACKNOWLEDGMENTS

This work is supported by a research grant from TÜBİTAK EEEAG with grant number 109E014.

## 7. REFERENCES

[1] W. Adams, G. Iyengar, C. Lin, M. Naphade, C. Neti, H. Nock, and J. Smith. Semantic indexing of multimedia content using visual, audio, and text cues. *EURASIP Journal on Applied Signal Processing*, 2:170–185, 2003.

[2] P. K. Atrey, M. A. Hossain, A. El-Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16:345–379, 2010.

[3] S. Ayache, G. Quénot, and J. Gensel. Classifier fusion for svm-based multimedia semantic indexing. *Advances in Information Retrieval*, pages 494–504, 2007.

[4] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *Multimedia, IEEE Transactions on*, 4(1):68–75, 2002.

[5] C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[6] L. Chaisorn, T. Chua, and C. Lee. A multi-modal approach to story segmentation for news video. *World Wide Web*, 6(2):187–208, 2003.

[7] C. Chang and C. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[8] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval: approaches and trends of the new age. In *MIR '05: Proc. of the 7th ACM SIGMM workshop on Multimedia information retrieval*, pages 253–262, New York, NY, USA, 2005. ACM Press.

[9] W. Hsu, L. Kennedy, C. Huang, S. Chang, C. Lin, and G. Iyengar. News video story segmentation using fusion of multi-level multi-modal features in trecvid 2003. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 3, pages iii–645. IEEE, 2004.

[10] G. Iyengar and H. Nock. Discriminative model fusion for semantic concept detection and annotation in video. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 255–258. ACM, 2003.

[11] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo. CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection. Technical report, Columbia University ADVENT #223-2008-1, August 2008.

[12] W. Lie and C. Su. News video classification based on multi-modal information fusion. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 1, pages I–1213. IEEE, 2005.

[13] L. Lin, G. Ravitz, M. Shyu, and S. Chen. Correlation-based video semantic concept detection using multiple correspondence analysis. In *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, pages 316–321. IEEE, 2008.

[14] P. Maragos, P. Gros, A. Katsamanis, and G. Papandreou. Cross-modal integration for performance improving in multimedia: A review. In P. Maragos, A. Potamianos, P. Gros, and B. Furht, editors, *Multimodal Processing and Interaction*, volume 33 of *Multimedia Systems and Applications Series*, pages 1–46. Springer US, 2008.

[15] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard. Yaafe, an easy to use and efficient audio feature extraction software. In *11th ISMIR conference, Utrecht, Netherlands*, 2010.

[16] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura. Video handling with music and speech detection. *Multimedia, IEEE*, 5(3):17–25, 1998.

[17] MPEG. Mpeg-7 reference software experimentation model, 2003. http://standards.iso.org/ittf/PubliclyAvailableStandards /c035364_ISO_IEC_15938-6(E)_Reference_Software.zip.

[18] A. Natsev, W. Jiang, M. Merler, J. Smith, J. Tesic, L. Xie, and R. Yan. Ibm research trecvid-2008 video retrieval system. In *Proc. of TRECVID*, volume 2008, 2008.

[19] P. Over, G. Awad, W. Kraaij, and A. F. Smeaton. Trecvid 2007–overview. In *TRECVID'07*, 2007.

[20] S. Ping and Y. Xiao-qing. Goal event detection in soccer videos using multi-clues detection rules. In *Management and Service Science, 2009. MASS'09. International Conference on*, pages 1–4. IEEE, 2009.

[21] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[22] D. Sadlier and N. O'Connor. Event detection in field sports video using audio-visual features and a support vector machine. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(10):1225–1233, 2005.

[23] C. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[24] C. Snoek, M. Worring, and A. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM, 2005.

[25] B. Truong and C. Dorai. Automatic genre identification for content-based video categorization. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 4, pages 230–233. IEEE, 2000.

[26] S. Tsekeridou and I. Pitas. Content-based video parsing and indexing based on audio-visual interaction. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(4):522–535, 2001.

[27] L. van der Maaten, E. Postma, and H. van den Herik. Matlab toolbox for dimensionality reduction. *MICC, Maastricht University*, 2007.

[28] Y. Wu, E. Chang, K. Chang, and J. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 572–579. ACM, 2004.

[29] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proc. of the 12th ACM Multimedia*, MULTIMEDIA '04, pages 572–579, New York, NY, USA, 2004. ACM.

[30] Z. Xiong. Audio-visual sports highlights extraction using coupled hidden markov models. *Pattern Analysis & Applications*, 8(1):62–71, 2005.

[31] D. Zhang and S. Chang. Event detection in baseball video using superimposed caption recognition. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 315–318. ACM, 2002.

[32] Q. Zhu, M. Yeh, and K. Cheng. Multimodal fusion using learned text concepts for image categorization. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 211–220. ACM, 2006.