FUSION METHODS FOR MULTI-MODAL INDEXING OF WEB DATA

Usman Niaz, Bernard Merialdo

EURECOM, Campus SophiaTech,
450 Route des Chappes, 06410 Biot FRANCE

ABSTRACT

Effective indexing of multimedia documents requires a multi-modal approach in which either the most appropriate modality is selected or different modalities are used in a collaborative fashion. A collaborative pattern is a model of combination between media that defines how and when to combine information coming from different media sources. Fusing information coming from different media seems a natural way to handle multimedia content. We focus on describing fusion strategies where the task is achieved through the use of different modalities. We browse through the literature looking at various state of the art multi-modal fusion techniques varying from naive combination of modalities to more complex methods of machine learning and discuss various issues faced with fusing several modalities having different properties in the context of semantic indexing.

1. INTRODUCTION

multi-modal fusion is gaining importance these days in the field of content analysis and retrieval as it benefits from diverse and complementary information from different media the content is represented with [1, 2]. This comes with a certain cost due to the higher complexity of multi-modal analysis as the modalities involved have different characteristics. There are various methods and different levels at which information from different media can be combined [1, 2, 3, 4]. The important questions thus raised when combining multi-modal information are when and how to fuse. The collaborative patterns answer precisely this question.

Collaboration of multi-modal information has been widely used for tasks including semantic indexing, copy detection in videos, video summarization, image and video search, object detection and recognition, biometrics etc.. Naturally each semantic concept or group of similar concepts exhibits different dynamics from others so their fusion pattern could be different also. This argument is strengthened by the fact that one classifier and one set of features does not perform the same for each concept [4] as some concepts are easy to detect with visual features only while others may not be. Fusion of different modalities is thus a sane way of performing analysis of multimedia web documents where each semantic concept exhibits a different pattern of combination. We start with differentiating the collaboration methodologies and then survey the state of the art multimedia fusion methods dividing them into early and late level of fusion.

Thanks to XYZ agency for funding.

2. COLLABORATIVE PATTERNS

Let us first define the terminology used in the paper. Let there be m modalities or sources of information like audio, video and text with the source i being represented with a d_i dimensional signature $\mathbf{x_i}$. The signature can be the low level feature vector $\mathbf{x_i} = \mathbf{F_i}$, where $\mathbf{F_i} = \{F_i^1, F_i^2, \ldots, F_i^{d_i}\}$ describes the modality or it can be the higher level decision of a classifier like score, probability etc.. In the later case $\mathbf{x_i} = D_i$ and the dimensionality d_i of the signature is 1 for one source. The nature of the signature mainly depends on the level at which information form the source is used.

A collaborative pattern is thus a fusion blackbox which takes all of these signatures as input,

$$\mathbf{v} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m}\}$$

and performs two tasks:

- Defines the method of combination of information from different sources. More specifically in the context of concept detection or semantic indexing [4] it is a function $f : R^d \rightarrow [0, 1]$ that takes the signature vector and outputs the probability of presence of the concept in the web document represented by that vector:

$$P(concept) = f(\mathbf{v})$$

- Learns a weight w_i for each source with $w_i \in \{0, 1\}$, to be used with the future instances of the source i. This weight controls the absence $w_i = 0$, presence $w_i = 1$ or importance of the modality i in the final output score.

2.1. Types of Collaborative Patterns

The methods of combination can be usually divided into two types according to [1] namely classification based and rule based methods. Classification based fusion can be done at

early and late level. A classifier can be built using the multi-modal feature vector that contains features from all the modalities. Discriminative classifiers like SVMs or Neural Networks can be trained if the dimension of the feature vectors remains the same. Generative learning like Bayesian Networks can also be used to learn from the multi-modal feature vectors. One advantage of generative classifiers is that they can also handle variable length vectors. Classifiers can also be used at the decision level to learn from the outputs of the classification results of the single media classifiers. In this case the scores from all the independent medium classifiers are combined in a single vector. These vectors are then learned by a secondary classifier for each concept. Regardless of the level at which the fusion is made the function $f(\mathbf{v})$ is represented in the simplest way as a classification function $f : R^d \rightarrow Y$, learning on the n training examples $\{(\mathbf{v_1}, y_1), \ldots, (\mathbf{v_n}, y_n)\}$ with $Y \in [0, 1]$.

Different rules can be employed to combine multi-modal information but it is more appropriate to do that at the decision level. A simple voting can be used to select the decision that occurs most often. Other methods like maximum or average of scores from classifiers of different modalities is also used in the literature. Linear weighted sum or linear weighted product can be used to combine the output scores. Rules can be custom defined based on the properties of the source or the classifier or some prior knowledge.

3. LEVELS OF FUSION

We borrow figure 1 from [1] to differentiate fusion strategies into early, late and hybrid of early and late methods. In figure 1 each Action Unit (AU) is usually a classifier that converts inputs into semantic level decisions while Feature Fusion (FF) and Decision Fusion (DF) units agglomerate features and decisions respectively. $D_{1,2}$ results from early fusion, $D_{m-1,m}$ comes from late fusion and D is the result of the hybrid fusion. Our survey of the multi-modal fusion methods in the context of content based indexing mainly divides the methods into early and late fusion strategies.

3.1. Early Fusion

In early fusion also referred to as fusion in feature space, unimodal features extracted from different data streams are integrated into the single large vector $\mathbf{v}$ for training. A certain pre-processing is performed like e.g. normalization so that features be on the same scale. Classifiers are trained for a semantic concept using these large multi-modal feature vectors.

Early fusion captures the true essence of multimedia collaboration as all the features are combined in a unified representation. There is only one learning phase handling all multi-modal features at once. If dimensions of the feature vectors from different media sources are fixed a discriminative classifier like SVM can be directly learned from the large feature

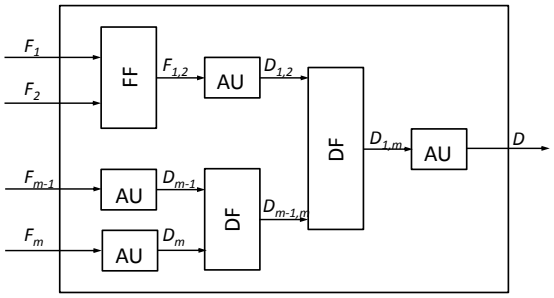

Fig. 1. Various levels of fusion

vector. Although fusion in feature space requires only one learning stage on the integrated feature vector it suffers from the problem of time synchronization and the unified representation of the media streams [1]. The choice of classifier is also tricky when fusing the multi-modal features. Some classifiers tend to work better for learning from visual data while others techniques may perform good on audio or textual features.

Guillaumin et al. [5] combine histogram based visual and binary text features in the Multiple Kernel Learning (MKL) framework to predict unlabeled examples for image classification. Law et al. [6] merge features early based on their statistics using a simple pattern of collaboration. The combine Bag of Words (BOW) feature with max pooling for discriminative image parts with average pooling for visual words from rather homogeneous image regions into a single normalized representation. The hybrid feature is trained using SVM for scene recognition. Kankanhalli et al. [7] have used early fusion of multi-modal features for face detections and traffic monitoring. They have used synchronized audio and visual streams along with textual metadata to describe a novel sampling based framework for multimedia analysis. They also select the most appropriate feature stream for the task by modeling a Markovian decision problem.

Snoek et al. [3] have performed early fusion by combining visual and textual descriptors in feature space. The visual vector contains pixel wise decision value for all the pixels in the best representative image segment which is combined with a histogram based textual feature extracted from the speech transcript. SVM classifiers are then trained for each semantic concept. Results yield a better performance for 6 out of 20 concepts using early fusion than late fusion.

Jiang et al. [8] ascertain the predominance of fusing audio and video features over single-modality based methods for multimedia event detection in videos. They build BOW representations for static visual, spatio-temporal video and MFCC audio features and learn non-linear SVMs for different categories. Zou and Bhanu [9] track humans in a cluttered environment by fusing multi-modal measurements. They exploit the correlation between the visual motion of the walking person with the corresponding step sound early at the feature level. Audio and visual features are combined and trained

using a Time Delay Neural Network (TDNN) where neurons in the hidden layers receive inputs from specific input nodes. They also train a Bayesian Network for the multi-modal features and show its superiority over the TDNN.

3.2. Late Fusion

Fusion can equally be performed at later stage integrating decisions of individual media classifiers thus called decision level fusion, also semantic level fusion [3]. Classifiers are learned for features of different media separately giving a decision like yes/no or in the form of a score or probability of presence of a semantic concept. The classification can be done in a way to obtain decisions having similar representation even if the features from different modalities have different representations. The independent decisions can be combined using different rules or classifiers can be built to learn from the output scores.

Fusion in decision space is easier to perform as it does not suffer from the representation problem early fusion does. The decisions from classifiers usually have similar format. Scores from new sources of information can be easily added to the final decision with only re-training the fusion part. Furthermore each different type of modality can have its own appropriate classifier. This can be considered slower compared to early fusion as modality dependent classifiers are first trained and their outputs are classified. Late fusion has been extensively used in the state of the art due its simplicity and scalability. We present some methods for late multi-modal fusion classified according to the methods of fusion used.

3.2.1. Classification based Late Fusion

Snoek et al. [3] have performed late fusion based on the outputs of the visual and textual analysis. Probabilistic output scores from visual and textual classifiers are combined in a single vector to form the multi-modal decision vector. These vectors are then trained with an SVM classifier to give the final output. From their results on 20 semantic concepts, 14 give better performance with late fusion. They conclude that the type of fusion ultimately depends on the type of concept chosen. Adams et al. [10] have also used decision level fusion for semantic indexing. multi-modal features are first extracted namely audio visual and text features. GMM, HMM and SVM classifiers are trained using these modalities and their decisions are joined to form a semantic feature. SVMs and Bayesian classifiers are then learned in the semantic space to obtain final decision for each concept.

Wang and Kankanhalli [11] exploit the correlation among different modalities and uncertainty of a classifiers for multimedia fusion using the portfolio theory. The portfolio theory is a theory of finance which attempts to maximize the returns on the collection of investments. More specifically a company is interested in making profitable investments on assets. Each investment has a certain level of risk or uncertainty and there is usually a correlation between investments i.e. collective investments may bare lower risk than individual investments. The portfolio theory finds the optimal portfolio that maximizes the expected return from the investments while minimizing the risk. The risk is modeled as variance of the classifier for each modality and the correlation among different modalities is calculated using Pearson's correlation coefficient. The classification performance or *expected return on investments* is maximized by finding ideal weights for each modality using the portfolio theory. Kim et al. [12] perform late fusion on classifier scores by learning in the maximal figure-of-merit (MFoM) for video retrieval. In the MFoM optimization framework weights for each score dimension are learned while maximizing the performance metric.

3.2.2. Rule based Late Fusion

McDonald et al. [13] use simple patterns for the collaboration of visual and text features for video retrieval. A video shot is represented using ASR text features and visual features such as HSV color, Canny edge and DCT based texture. Output probability scores are obtained using language modeling approaches separately for text and visual features and are combined using different weights. The superiority of fusion based methods over single modality classifiers is verified with different combination methods. Yan et al. [14] combine retrieval results from multiple modalities for video retrieval. They use EM algorithm to learn linear combination of weights of the scores of classifiers based on text (ASR, closed captions and video OCR) and visual (color and edge histogram) features.

Strat et al. [4] present three custom defined rules to combine decisions of dozens of multi-modal classifiers for video concept detection. For each concept each classifier provides a ranked list of video shots indicating the likelihood of presence of the concept. The first technique clusters the classifier outputs manually based on their nature and principle of operation like the type of descriptor, the machine learning algorithm used, the type of modality etc.. Scores inside each cluster are first fused together using weighted arithmetic mean and then the clusters are fused similarly. They conclude that going from most similar to the most different to cluster descriptors and classifiers was a good strategy.

Their second approach performs an agglomerative clustering of scores from highly correlated sources while discarding irrelevant classifiers. First the relevance of each classifier is judged and those with near-random performance are filtered out. Correlation between pairs of classifiers is measured and the two most highly correlated are fused into one classifier. This process is repeated until no sufficiently correlated pair is left, resulting in a bunch of fused classifiers. These classifiers are fused using weighted fusion with individual relevance as weights. They are also fused separately using a neighborhood based method after dimensionality reduction using PCA. For

this each test shot is considered positive if no shot from training set is present around in the neighborhood of radius d on each of the reduced dimensions. This strategy is applied because positive shots are assumed to be very low in number, rare and different from negative shots. In their third approach they use a community detection approach to discover the clusters while the fusion is done similarly. Results indicate manually selecting clusters performs best.

Hua et al. [15] obtain a better decision by combining a set of decisions obtained from different data sources to perform image retrieval. The decisions are fused based on certain cues obtained from the degree of attention people pay to certain objects, e.g. the strength of a sound, the speed of a motion, the size of an object, and so on. The cues are measured using distance between two images from features like color histogram, color moment, wavelet, block wavelet, correlogram, blocked 5 correlogram. The similarity scores between two images for each feature type are fused using an attention fusion function which is a variation of linear weighted sum.

Bredin et al. [16] divides a video in segments and selects the maximum score from different classifiers on multiple modalities as the best decision for that segment for identifying person P in that video segment. Liu et al. [17] present the Selective Weighted Late Fusion (SWLF) scheme to avoid overfitting where adding scores from more classifiers decrease performance of the final fused output. For each concept SWLF sorts scores from different multi-modal classifiers and selects top N linearly if the next one improves performance.

4. ISSUES OF COLLABORATIVE FUSION

The correlation between modalities, calculated at feature level or semantic level, should be used to find helpful cues to effective patterns of collaboration [1]. The question of selecting the most proper modality for a concept can be answered automatically by a classifier or through linear weighted combination, or can be manually defined based on some prior knowledge. Due to the difference in the sampling rate and processing time of different modalities the synchronization of information is an important issue to consider. Authors in [16] tackle this issue by aligning various multi-modal data streams into finest common segmentations and take the final decision at this segmentation level. However this temporal partitioning may not always prove helpful [8] and the hybrid features could be trained once for a video clip using early fusion.

5. CONCLUSIONS

State of the art shows that fusing different modalities outperforms the overall best performance from the single best classifier for each concept as even the weak performing ones bring complimentary information. Several successful attempts have been made to understand the correlation between different modalities and use it to help the video analysis. Furthermore importance of a modality or a classifier of certain modality can be calculated. All these parameters help finding the optimal collaborative pattern of fusion where different modalities complement each other for video analysis.

6. REFERENCES

[1] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[2] C. G.M. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, pp. 5–35, 2003.

[3] C. G. M. Snoek, M. Worring, and A. Smeulders, "Early versus late fusion in semantic video analysis," 2005, ACM MM.

[4] S. T. Strat, A. Benoit, H. Bredin, G. Quénot, and P. Lambert, "Hierarchical late fusion for concept detection in videos," 2012, ECCV.

[5] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *CVPR*, 2010.

[6] M. Law, N. Thome, and M. Cord, "Hybrid pooling fusion in the bow pipeline," 2012, ECCV.

[7] M. S. Kankanhalli, J. Wang, and R. Jain, "Experiential sampling in multimedia systems," *Trans. Multi.*, vol. 8, no. 5, pp. 937–946, Oct. 2006.

[8] Y. G. Jiang, G. Ye, S. F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," 2011, ICMR.

[9] X. Zou and B. Bhanu, "Tracking humans using multi-modal fusion," 2005, CVPR.

[10] W. H. Adams, G. Iyengar, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith, "Semantic indexing of multimedia content using visual, audio and text cues," *EURASIP*, vol. 2, pp. 170–185, 2003.

[11] X. Wang and M. Kankanhalli, "Portfolio theory of multimedia fusion," 2010, MM.

[12] I. Kim, S. Oh, B. Byun, A. G. A. Perera, and C. H. Lee, "Explicit performance metric optimization for fusion-based video retrieval," 2012, ECCV2.

[13] K. McDonald and A. F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," 2005, CIVR.

[14] R. Yan, J.n Yang, and A. G. Hauptmann, "Learning query-class dependent weights in automatic video retrieval," 2004, MULTIMEDIA '04, pp. 548–555, ACM.

[15] X. S. Hua and H. J. Zhang, "An attention-based decision fusion scheme for multimedia information retrieval," 2004, PRCM.

[16] H. Bredin, J. Poignant, M. Tapaswi, G. Fortier, V. Le, T. Napoleon, H. Gao, C. Barras, S. Rosset, L. Besacier, J. Verbeek, G. Quénot, F. Jurie, and H. Ekenel, "Fusion of speech, faces and text for person identification in tv broadcast," 2012, ECCV.

[17] N. Liu, E. Dellandrea, C. Zhu, C. E. Bichot, and L. Chen, "A selective weighted late fusion for visual concept recognition," 2012, ECCV.