Multimedia Evidence Fusion for Video Concept Detection via OWA Operator

Ming Li[1, 2, 3] Yan-Tao Zheng[2] Shou-Xun Lin[1] Yong-Dong Zhang[1] Tat-Seng Chua[2]

[1]Key Laboratory of Intelligent Information Processing, ICT, CAS, Beijing, China 100190
[2]Department of Computer Science, National University of Singapore, Singapore 117543
[3]Graduate School of the Chinese Academy of Sciences, Beijing, China 100039
[1]{mli, sxlin, zhyd}@ict.ac.cn, [2]{yantaozheng, chuats}@comp.nus.edu.sg

Abstract. We present a novel multi-modal evidence fusion method for high-level feature (HLF) detection in videos. The uni-modal features, such as color histogram, transcript texts, etc, tend to capture different aspects of HLFs and hence share complementariness and redundancy in modeling the contents of such HLFs. We argue that such inter-relation are key to effective multi-modal fusion. Here, we formulate the fusion as a multi-criteria group decision making task, in which the uni-modal detectors are coordinated for a consensus final detection decision, based on their inter-relations. Specifically, we mine the complementariness and redundancy inter-relation of uni-modal detectors using the Ordered Weighted Average (OWA) operator. The 'or-ness' measure in OWA models the inter-relation of uni-modal detectors as combination of pure complementariness and pure redundancy. The resulting weights of OWA can then yield a consensus fusion, by optimally leveraging the decisions of uni-modal detectors. The experiments on TRECVID 07 dataset show that the proposed OWA aggregation operator can significantly outperform other fusion methods, by achieving a state-of-art MAP of 0.132.

Keywords: OWA, Fusion, Video Concept Detection.

1 Introduction

The multi-modal and multimedia nature of video demands the judicious use of all information channels, such as visual, textual, acoustic, etc, to analyze its semantic content [1] [2] [3] [13]. It is intuitive that the use of more information channels tends to deliver better performance than a single information channel. In the task of high-level feature (HLF) extraction in video, most state-of-arts systems [4] [5] adopt a multi-modal multi-feature framework to exploit multimedia evidences from different channels. The widely used multi-modal evidence fusion schemes can be generally classified into 2 types: early-fusion and late-fusion.

Early-fusion is the scheme that integrates uni-modal features before learning the video concepts. Though early-fusion is straightforward, it usually does not work well. This is so because the features of different modalities are not isomorphic. They tend to possess different dimensionality, representation form and measurement space. It is

thus not reasonable to simply concatenate them into a common feature space. Moreover, the higher dimensionality caused by concatenation will worsen the scarceness of training data and curse of dimensionality. In fact, the early-fusion is more suitable for fusing raw data within single type of feature, such as to capture spatial information with different granularity, i.e. fusing grid-based visual features and image-based visual feature of same type.

In contrast to early-fusion, late-fusion first reduces uni-modal features to separately learned concept detection decisions and then integrates them to the final detection decision [6]. The late-fusion has been reported to outperform early fusions significantly [6]. The major criticism for late-fusion is its loss of linear correlation of multi-modal features. However, when the features come from completely different modalities or channels, their linear correlation becomes less critical. Instead, the complementariness and redundancy of individual uni-modal detectors are the key to effective fusion.

Hence, we need to develop a principled method to fuse the outputs of uni-modal detectors to form a final consensus detection decision. From this perspective, the multi-modal fusion can be regarded as a multi-criterion group decision making problem (GMD) [7], in which the group of uni-modal detectors should be consolidated synergically towards the common task. We therefore formulate the multi-modal fusion as an information aggregation task in the framework of group decision making (GMD) problem. Specifically, we employ the Ordered Weighted Average (OWA) operator to aggregate the group of decisions by uni-modal detectors, as it has been reported to be an effective solution for GMD problem [10]. Compared to existing fusion methods, the major advantage of OWA is that its 'or-ness' measure can explicitly reflect the complementariness and redundancy interrelation of uni-modal detectors, in between either a pure 'and-like' or a pure 'or-like' manner. Our main contributions are twofold: (a) we formulate the multi-modal fusion as a group decision making problem (GMD); (b) we employ the OWA operators to consolidate the uni-modal detections by mining their complementariness and redundancy inter-relations. Experiments on TRECVID 07 dataset show that the proposed approach can significantly outperform other fusion methods.

2 Related Work

Our proposed fusion method belongs to late-fusion scheme, as it manipulates the outputs of uni-modal detectors. There are mainly two types of late-fusion schemes. The first views the late-fusion problem from the perspective of statistical machine learning. It regards the learned outputs of uni-modal detectors as input for another layer of statistical learning. For example, Snoek et al. [6] employed SVM to fuse the output of individual classifier on each set of uni-modal features. Though it has been reported to deliver outstanding performance [6], it has the following disadvantages: (a) it often requires extensive parameter tuning and model selection for the additional round of statistical machine learning; and (b) the possibility of over fitting in the supervised learning process limits its generalization ability.

The second type considers late-fusion as an information aggregation task, from the perspective of Information Theory. The fusion now aims to maximize the aggregated information by assigning proper weights to individual information channels. The examples include heuristic fusions like Weighted Average and Adaboost; and non-heuristic fusions like Maximum (Max), Minimum (Min) and Average [12] scheme. The major drawbacks of non-heuristic fusion methods are that they are not adaptive to different aggregation problems; and they may not converge well or the converged result may be local rather than global optimum.

Similar to the aforementioned fusion schemes, our proposed fusion belongs to the second type. However, different from the fusion methods above, our proposed OWA based fusion provides a general approach of parameterized decision aggregation, with an efficient global optimal weight vector searching strategy. To our best knowledge, this is also the first approach to investigate the inter-relations of uni-modal detectors, from the perspective of Information Theory.

3 OWA-Based Multi-Model Fusion

For each shot, a set of uni-modal features $\{\mathbf{x}_i\}_{i=1}^n$ are first extracted from *n* different modalities $\{\mathbf{M}_i\}_{i=1}^n$, such as visual, auditory and text. The decision function T, such as Support Vector Machine, is then applied on each uni-modal to yield a separately learned concept score $a_i = T(\mathbf{x}_i)$, and $a_i \in [0,1]$. Our target now is to apply the OWA operator to learn the complementariness and redundancy of the group of uni-modal detection scores $\{a_i\}_{i=1}^n$ to yield a consensus final detection score.

The inter-relation of uni-modal detectors has two extreme cases: (a) complete redundancy; and (b) complete complementariness. In the case of complete redundancy, all uni-modal detectors contribute similarly on the given HLF detection task. This implies the final detection decision is positive, only if all the detectors deliver positive decisions. In this scenario, these detectors actually share an and-like inter-relation. In the case of complete complementariness, the uni-modal detectors perform differently on the given detection task. This means that the final decision can be positive, if one of the detectors yields positive decisions. In this case, the detectors share an or-like inter-relation. Our argument is that the inter-relation of detectors lies somewhere in between these two extremes. The OWA operator is exploited to discover

3.1 OWA Aggregation Operator

An OWA operator is a mapping $F: \Re^n \to \Re$ from *n* uni-modal detection decisions to one final detection decision with a weight vector $W = (w_1, \cdots, w_n)^T$ subjects to

$$w_1 + \cdots + w_n = 1, 0 \le w_i \le 1, i = 1, \cdots, n \quad \textbf{(1)}$$

and such that

$$F(a_1,\ldots,a_n)=\sum_{i=1}^{n} w_i b_i \quad (2)$$

where a_i is the decision from uni-modal feature $\mathbf{x}_i$ and b_i is the i^{th} largest element of the aggregated decision $\{a_1,\cdots,a_n\}$. Note that the re-ordering step is a critical aspect of OWA operator, in which a_i is not associated with a particular weight w_i, but rather a weight w_i is associated with a particular ordered position b_i of a_i.

Yager [8] introduced two characterizing measures associated with the weight vector *W* of an OWA operator. The first one is the measure of 'or-ness' of the aggregation operator and is defined as:

$$orness(W)=\frac{1}{n-1}\sum_{i=1}^{n}(n-i)w_i \quad (3)$$

The *orness* measures how much the aggregation associated with vector *W* is like "OR" aggregation operator. When $W=[1\,0\,0\,...]$, then $orness(W)=1$ and the final decision $F(a_1,\ldots,a_n)=\text{MAX}(a_i)$. This shows a complete redundancy inter-relation of uni-modal detectors, as the fusion only takes into account the decision of largest value. In contrast, when $W=[0\,0\,...1]$, $orness(W)=0$ and the final decision $F(a_1,\ldots,a_n)=\text{MIN}(a_i)$. This shows a complete redundancy inter-relation, as the fusion only takes into account the uni-modal decision of small value. Intuitively, Max is a pure "OR" operator and Min is a pure "AND", which are the two extreme conditions of information aggregation [8]. Obviously Max, Min and Average are special cases of OWA aggregation operator with certain weight vector.

The second measure in OWA is the 'dispersion' (entropy) of the aggregation operator. It is defined as:

$$dispersion(W)=-\sum_{i=1}^{n} w_i \ln w_i \quad (4)$$

It measures how much information is taken into account in the aggregation. We can also derive the dispersions of Max, Min and Average aggregation operators; i.e. 0, 0 and $\ln n$, respectively. The dispersion is actually a Shannon entropy of weights. In a certain sense, the more disperse W is, the more information about the individual detection decision is being used in the aggregation of OWA operator [8].

4 Optimal OWA Weight Learning

4.1 Orness-Dispersion Space (OD-space)

Intuitively, the optimal weights can be obtained by performing grid search in the *n*-dimension weight vector space, where *n* is the number of uni-modal concept detection

decisions to aggregate. Such simple grid search is, however, proved to be computationally intractable, due to the high dimensionality of weight vector space [8].

Fortunately, it is found that the weight vectors with similar Orness and dispersion values are of similar aggregating performance [8]. Thus we can transform the search space from *n*-dimensional weight vector space into a 2 dimensional OWA OD-space. Given

$$orness(W) = \alpha \ , \quad dispersion(W) = \beta \tag{5}$$

To derive W subjects to Eqn (5) is equivalent to solving the following problem:

$$\begin{gathered} \frac{1}{n-1}\sum_{i=1}^{n}(n-i)w_i = \alpha, 0 \le \alpha \le 1 \\ -\sum_{i=1}^{n} w_i \ln w_i = \beta, 0 < \beta < \ln n \\ \sum_{i=1}^{n} w_i = 1, 0 \le w_i \le 1 \end{gathered} \tag{6}$$

4.2 Orness/ Dispersion Max-space

As Eqn(6) cannot be solved analytically, we exploit some approximation to transform it into a optimization problem, which has many existing solutions, such as Lagrange multipliers. Here, we consider mainly two kinds of optimization formulation to solve for weights *W*.

First, we can optimize the variability by maximizing the dispersion or minimizing the variance of the weights while keeping the Orness at a fixed level [9].

$$\begin{gathered} \text{Maximize} \quad \frac{1}{n-1}\sum_{i=1}^{n}(n-i)w_i \\ \left(\text{or Minimize} \quad Var(W) = \frac{1}{n}\sum_{i=1}^{n} w_i^{\,2} - \frac{1}{n^2}\right) \\ \text{subject to} \quad -\sum_{i=1}^{n} w_i \ln w_i = \beta, 0 < \beta < \ln n \ ; \quad \sum_{i=1}^{n} w_i = 1, 0 \le w_i \le 1 \end{gathered} \tag{7}$$

The second solution is to maximize the Orness, while keeping the dispersion at a fixed level [10].

$$\begin{gathered} \text{Maximize} \quad -\sum_{i=1}^{n} w_i \ln w_i \\ \text{subject to} \quad \frac{1}{n-1}\sum_{i=1}^{n}(n-i)w_i = \alpha, 0 \le \alpha \le 1 \\ \sum_{i=1}^{n} w_i = 1, 0 \le w_i \le 1 \end{gathered} \tag{8}$$

Both optimization formulations above are trying to transform the 2-dimension OD-space into 1-dimension maximum line space (Max-space). Although the mapping is

not a full cover of the OD-space, Max-spaces are the most significant sub-spaces of it. In Orness Max-space modeled by Eqn (7), the weights of different dispersion are given with maximal Orness which means the operator is as disjunctive as possible using at least certain percentage of available information [9]. In Dispersion Max-space represented by Eqn (8), the weights of different Orness are given with maximal dispersion which means most individual criteria are being used in the aggregation that gives more robustness [10]. Both Eqn (7) and Eqn (8) are optimal problems which can be analytically solved using Lagrange multipliers. The results are optimal (maximal or minimal) Orness (or dispersion) weight vectors of different dispersions (or Orness). The weight vectors with different Orness and dispersion are evaluated in the cross-validation set and the one that gives the highest precision is chosen to be the OWA aggregation operator's weight vector.

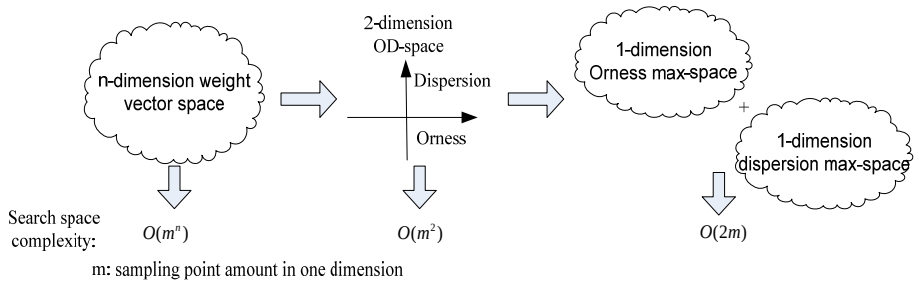

Figure 1. Search space complexity comparison among different spaces

4.3 Theoretic Analysis of OWA

Figure 1 compares the search space complexity among different weight vector search spaces. As shown, by transforming the search for optimal weights into an optimization task, the complexity of OWA learning is greatly reduced from n-dimension weight vector space $O(m^n)$ to Orness/dispersion max-space $O(2m)$ where m is the number of sampling points in one dimension. Besides, given a fixed Orness (dispersion), the first component of OWA weights vector can be calculated by solving an equation and others can be simply generated from the first component [10]. The complexity of OWA learning is linear to the training dataset, while SVM based fusion has a learning complexity of $O(N^3)$, due to the convex quadratic programming (QP) formulation of SVM. The Max, Min and Average are non-heuristic fusions that don't need training process. However, their performance is worse than that of the OWA fusion, which will be shown in next section.

5 Experiment and Results

5.1 Testing Dataset and Experimental Setup

We evaluate the proposed OWA-based fusion on TRECVID 07 dataset [11]. The TRECVID 2007 dataset comprises 100 hours of documentary video with ~40, 000 key frames. The dataset is equally divided into development and test sets, with each containing approximately 50 hours and around 20,000 key frames. In our experiments, we use the 20 semantic concepts listed in Figure 2 used in the TRECVID 07 valuation [11].

We further split the development set into 70% for training and 30% for validation. The training set is used to train the SVM detectors for individual uni-modal features; while the validation set is used to train the various fusion methods. For simplicity, we use a keyframe to represent a shot and extract 6 types of features from the keyframe of each shot. They are color correlogram (CC), color histogram (CH), color moment (CM), edge histogram (EH), texture co-occurrence (TC) and wavelet texture (WT). For CM, we compute the first 3 moments of RGB color channels over 5 $\times$5 grids to form a 225D feature vector. For WT, we divide a key frame into 4 $\times$ 3 grids and compute the variance in 9 Haar wavelet sub-bands for each grid. This gives rise to a 108D feature vector for a keyframe. The evaluation criteria is the mean average precision (MAP), which is the mean of average precision (AP) of each concept.

Table 1. MAPs of each uni-modal feature set

Feature	CC	CH	CM	EH	TC	WT
MAP	0.082	0.065	0.084	0.082	0.028	0.033

Table 2. MAP of various fusion models

Fusion	OWA	SVM	Adaboost	Max (OR)	Min (AND)	Average
MAP	0.132	0.122	0.064	0.088	0.075	0.12

5.2 Experiments and Discussion

We first perform SVM-based concept detection on each uni-modal feature. The cost and gamma parameters of SVM are set, based on cross validation. Table 1 shows the MAP of each feature set. Based on the individual concept scores of uni-modal detector, we apply the OWA operator to learn the optimal weights for ordered concept scores. As shown in Table 2, OWA achieves the best MAP of 0.132, which is substantially higher than that of all uni-modal detectors. This demonstrates that different features do complement each other and the proper fusion of decisions of all uni-modal concept detectors can yield a much more satisfactory detection performance.

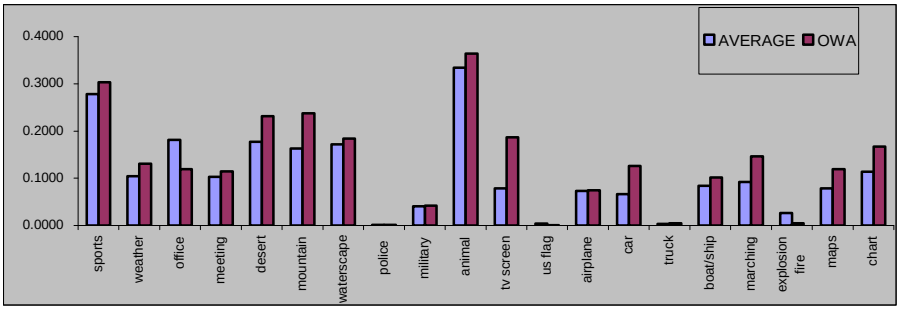

Figure 2. MAP of each concepts by OWA and Average fusion respectively.

Next, we compare OWA-based fusion with 5 commonly used late-fusion methods: SVM, Adaboost, Max, Min and Average. The SVM-based fusion is to take uni-modal detection score as input and yield a final detection decision in a supervised learning process. The Max fusion is a specialization of OWA fusion. It takes the maximum detection score of all uni-modal features as the final detection score, which is effectively a pure OR operator on the group of uni-modal detection decisions. The Min fusion is also a specialization of OWA fusion. It takes the minimum detection score of all uni-modal features as the final detection score, which is effectively a pure AND function. Average fusion is to take the mean of all uni-modal detection scores as the final detection score.

Table 2 lists the MAP of these 5 fusion methods. As shown, the Max and Min fusions do not yield good detection performance over uni-modal detectors. This is so because these two fusion schemes over-simplify the inter-relation among different features as either complete complementariness (pure OR inter-relation) or complete redundancy (pure AND inter-relation). On the other hand, the Average fusion gives a better MAP of 0.12, which demonstrates that the inter-relation between different features lies between the pure complementariness and pure redundancy. That the best MAP is achieved with OWA fusion, which demonstrates that OWA is effective to learn the inter-relations of different uni-modal features. We also observe that the SVM based fusion achieves an MAP of 0.122, which is much lower in performance than that of OWA but comparable to Average fusion. However, considering the expensive learning process of SVM with little improvement in performance over the simple Average fusion, the SVM based fusion is not cost-effective.

In order to capture a clearer view on the performance of OWA fusion, we examine the fusion performance on each concept. Figure 2, shows the MAP of Average and OWA fusion for each concept. We observe that OWA outperforms the Average fusion in 17 out of 20 concepts. Of the 3 concepts that Average fusion outperforming OWA fusion, we found that the number of positive training samples in these concept classes are too small. These 3 concepts have an average of 613 positive samples, while the average number of samples for all concepts is 1591. This is one reason why OWA, which is essentially a supervised learning process, might yield inaccurate results.

Table 3. MAP of various fusion models

	OWA	**[4]**	**[5]**	**[12]**
MAP	0.132	0.1311	0.099	0.098

Benchmark on TRECVID 07: We compare our system with other reported systems. Table 3 tabulates the MAPs of the top performing systems in TRECVID 2007. As shown, the proposed approach based on OWA outperforms most of the existing systems, and delivers a comparable result with the best reported system [4], which however exploited a computationally expensive Multi-Label Multi-Feature learning process [4].

6 Conclusion And Future Work

We have proposed a multi-modal fusion method for HLF detection in video. By exploiting the OWA operator, we mine the inter-relation of uni-modal detectors and then coordinate them with averaging weights to yield a final consensus detection decision. The experiments on TRECVID 07 showed that the OWA based fusion can outperform other fusion method, such as SVM, Adaboost, with statistically significant improvements.

Several issues remain open. First, we do not exploit other subspaces, such as Variance Min-space, for weight learning. Though this does not improve learning efficiency, it does give a different perspective to learn OWA weights. Second, the Bayesian formulation can be incorporated to learn the weights in a probabilistic manner.

Acknowledgments. This work was supported in part by the National Nature Science Foundation of China (60873165), the National Basic Research Program of China (973 Program, 2007CB311100), the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416).

References

1. Chang, S.-F., Hsu, W., Kennedy, L., Xie, L., Yanagawa, A., Zavesky, E., and Zhang., D.-Q. 2006. Columbia university trecvid-2005 video search and high-level feature extraction. In TREC Video Retrieval Evaluation Proceedings, March 2006.
2. Dorai, C., and Venkatesh., S. 2003. Bridging the semantic gap with computational media aesthetics. In IEEE MultiMedia, 10(2):15-17, 2003.
3. Hauptmann, A. G., Chen, M.-Y., Christel, M., Lin, W.-H., Yan, R., and Yang, J. 2006. Multi-lingual broadcast news retrieval. In Proceedings of TREC Video Retrieval Evaluation Proceedings, March 2006.

4. Mei, T., Hua, X., Lai, W., Yang, L., Zha, Z., Liu, Y., Gu, Z., Qi, G., Wang, M., Tang, J., Yuan, X., Lu Z., and Liu, J. 2007 MSRA-USTC-SJTU at TRECVID 2007: High-level feature extraction and search. In http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html
5. Le, HD., Satoh S., and Matsui, T. 2007. NII-ISM, Japan at TRECVID 2007: High Level Feature Extraction. In http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html
6. Snoek, C., Worring, M., Gemert, J., Geusebroek, J.-M., and Smeulders, A. 2006. The challenge problem for automated detection of 101 semantic concepts in multimedia. In Proceedings of ACM MM, pages 421-430. 2006
7. Kacprzyk, J., Fedrizzi, M., and Nurmi, H. 1997. OWA operators in group decision making and consensus reaching under fuzzy preferences and fuzzy majority. In: Yager, R.R., Kacprzyk, J. (Eds.), The Ordered Weighted Averaging Operators: Theory and Applications, Kluwer Academic Publishers, pages 193-206.
8. Yager, R.R. 1988. Ordered weighted averaging aggregation operators in multi-criteria decision making. In IEEE Tran. On Systems, Man and Cybernetics, 18(1988) pages 183-190
9. Marchant, T. 2006. Maximal orness weights with a fixed variability for OWA operators. In International Journal of Uncertainty Fuzziness and Knowledge Based Systems, 14(2006) pages 271-276
10. Fuller, R., and Majlender, P. 2001. An analytic approach for obtaining maximal entropy OWA operator weights. In Fuzzy Sets and System, 124(2001) pages 53-57
11. Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval MIR '06. ACM Press, pages 321-330.
12. Ngo, C., Jiang, Y., Wei, X., Wang, F., Zhao, W., Tan, H., and Wu, X. 2007. Experimenting vireo-374: Bag-of-visual-words and visual-based ontology for semantic video indexing and search. In TREC Video Retrieval Evaluation Proceedings, Nov 2007.
13. Magalhães, J. and Rüger, S. 2007. Information-theoretic semantic multimedia indexing. In Proceedings of the 6th ACM international conference on Image and video retrieval (CIVR '07), July 2007.