Building Multi-model Collaboration in Detecting Multimedia Semantic Concepts (Invited Paper)

Hsin-Yu Ha, Fausto C. Fleites, Shu-Ching Chen
School of Computing and Information Sciences,
Florida International University,
Miami, FL 33199, USA
{hha001,fflei001,chens}@cs.fiu.edu

***Abstract*—The booming multimedia technology is incurring a thriving multi-media data propagation. As multimedia data have become more essential, taking over a major potion of the content processed by many applications, it is important to leverage data mining methods to associate the low-level features extracted from multimedia data to high-level semantic concepts. In order to bridge the semantic gap, researchers have investigated the correlation among multiple modalities involved in multimedia data to effectively detect semantic concepts. It has been shown that multimodal fusion plays an important role in elevating the performance of both multimedia content-based retrieval and semantic concepts detection. In this paper, we propose a novel cluster-based ARC fusion method to thoroughly explore the correlation among multiple modalities and classification models. After combining features from multiple modalities, each classification model is built on one feature cluster, which is generated from our previous work FCC-MMF. The correlation between medoid of a feature cluster and a semantic concept is introduced to identify the capability of a classification model. It is further applied with the logistic regression method to refine ARC fusion method proposed in our previous work for semantic concept detection. Several experiments are conducted to compare the proposed method with other related works and the proposed method has outperform other works with higher Mean Average Precision (MAP).**

Keywords*—Semantic concept detection, Multi-model Fusion, Feature Correlation

I. Introduction

Recently, the amount of multimedia data has been drastically increasing as the development of all kinds of handheld device, such as smart phones and digital cameras, will allow people easily share their life by uploading multimedia data in one-click. Take facebook as an example, at the beginning of 2013, they announced their one billion users have uploaded 240 billion photos since the site's launch ,and each user would have more than 200 photos uploaded on an average basis. How to effectively analyze multimedia data started to draw research attention a couple decades ago, and it has become more and more crucial as a result of increasing a variety of multimedia data and enormous amount of data are populating around the world. In addition, serious challenges have been blocking the way for better multimedia data management and efficient data retrieval, such as a great diversity of data representation, high computational complexity, and lack of strong computing power.

To overcome the obstacles to multimedia research, some researchers tried to make progress by utilizing highly discriminative and robust features [1] such as Scale Invariable Feature Transformation (SIFT) [2][3] and Histogram of Oriented Gradients (HOG) [4][5]. Considering only single modality, such as analyzing audio signal for automatic transcription of speech, leveraging color features for scene recognition or using temporal features to detect different action, has also been greatly investigated. However, it has shown significant limitations while coping with tasks, which have multiple modalities involved, for instance, multimedia retrieval and multimedia event detection.

Exploiting information extracted from all the involving multiple modalities has been proven to be advantageous to multimedia analysis. Nonetheless, several major issues have not yet been adequately addressed. As a starter, handling data with different presentations such as visual, audio and text, is an issue. Moreover, how to fully employ all the given information, such as the correlation among different modalities, is also quite challenging. The other interesting topics would be identifying the useful modalities or classification models and fusing them to strengthen the achievement of multimedia related tasks.

In this paper, built on our previous works [6][7], we proposed a Multi-Model Collaboration (MMC) framework for multimedia semantic concept detection by introducing a novel cluster-based ARC fusion method, where ARC stands for adjustment, reliability and correlation of the intervals to the target semantic concept. In our previous works, features extracted from multiple modalities are transformed into feature clusters with high intra-correlation and low inter-correlation. The proposed cluster-based ARC fusion method is improved by considering the correlation between transformed feature clusters and the target concepts. Logistic regression is also applied to optimize the ranking scores in the fusion process. Finally, a threshold is set up based on experiments to eliminate the unproductive classification models build on feature clusters, hence only the classification models which have higher reliability are deployed in the fusion process.

The rest of the paper is organized as follows. Related work is introduced in Section 2. Section 3 presents the overview of the proposed MMC framework. Section 4 describes the experimental results and the framework evaluation. At the end, section 5 concludes this paper.

COLLABORATECOM 2013, October 20-23, Austin, United States

DOI 10.4108/icst.collaboratecom.2013.254110

II. Related Work

In the multimedia research domain, multi-modal fusion has attracted much attention not only because uni-modal approaches have their limitation to achieve complicated tasks but also because multi-modal approaches provide resourceful information for various multimedia analysis tasks. Researchers who have participated in significant image retrieval tasks, e.g., ImageCLEF [8] and TRECVid [9], have witnessed how multi-modal fusion takes over the major role in multimedia analysis. The organizers of ImageCLEF have been providing multimedia databases including images with associated text since 2003 for participants to investigate the effectiveness of multi-modal retrieval [10]. TRECVid, which has involved over 1,200 researchers from hundreds of research groups around the world, has been holding a benchmark annual activity to encourage researchers addressing multimedia related tasks, specifically semantic concept detection from video is one of the major tasks involving multiple modalities [9].

Based on a comprehensive survey article about multi-modal fusion, the fusion strategies can be mainly categorized into early and late fusion methods [11]. Early fusion can be referred to as an integration of features extracted from multiple modalities; on the other hand, an integration of the intermediate results is referred to as late fusion. We will briefly go over the related works and distinguish our proposed work from them.

With regards to early fusion, the basic approach simply concatenates features from multiple modalities into one large feature set and converts it into one consistent representation [12] [13][14]. Given one complete feature set, several research works applied Canonical Correlation Analysis (CCA) to model the correlations between features [15][16][17]. Sargin et al. [15] applied CCA to fuse audio and lip texture features to achieve audiovisual synchronization. Liu et al. [16] proposed a audio-visual fusion framework, in which CCA is used to project the audio and visual features into more compact subspaces. Hence, the correlation conveyed in the original audio and visual feature space can be preserved; meanwhile, model efficiency can be improved in the more compact feature spaces. Different from these related works, instead of leveraging correlation among features, the proposed framework increases the granularity of correlation to explore the correlation within feature-value pairs and better build the classification models on the finer captured correlations.

Late fusion, also called decision-level fusion, integrates the classification results from different modalities and generates only one result[18][19][20]. Usually, each modality is analyzed independently, so it has the flexibility to select the most suitable approach for different modalities, such as latent semantic Index (LSI) for textual modality, and hidden markov model (HMM) for audio or video modality. In addition, since the classification results collected from multiple modalities usually have the same representation, it is easier to fuse the results. However, each modality usually generates its own decision result independently, and the correlation among different modalities are overlooked in many related works adopting late fusion strategy. For example, Potamianos et al. [21] combined classification results from audio modality and visual modality and fuse two independent results with a linear weighted sum method. Chen et al. [7] proposed a fusion method called ARC and it's goal is to achieve a performance gained from all individual models. Inspired by ARC, the proposed cluster-based ARC includes all the possible correlations among modalities at feature-value pair, feature, and classification model levels to refine the factor used in model fusion and enhance the precision of semantic concept detection.

III. Overview of MMC

MMC is a multiple-model collaboration framework designed and implemented for multimedia semantic concept detection. It is improved from our previous works by introducing a novel fusion method named cluster-based ARC. With the enhanced fusion method, the correlation between classification models, which are built up from all the features extracted from multiple modalities, and the target concepts is thoroughly explored. Fig. 1 and Fig. 3 depict the training and testing components of the proposed framework, respectively.

A. *Training section of MMC framework*

In the training section, the first three steps, e.g., pre-processing, feature-value pair projection, and feature-value pair clustering, are the same processes as proposed in our previous work FCC-MMF [6]. In the next highlighted green square, a new factor γ is introduced to refine ARC fusion method proposed in [7]. Logistic regression is applied to obtain the optimal weighting factor in integrating γ and α. Finally, feature cluster selection is performed based on model efficiency.

From the beginning, features are extracted from all the involving modalities so that information containing different characteristics can be fully exploited. The pre-processing step includes redundant text removal, discretization and normalization to formalize the feature presentation. Multiple correspondence analysis (MCA) is employed to analyze the correlation among all the feature-value pairs by projecting each feature-value pair as one point onto the two major principal components. Subsequently, because K-medoids algorithm is one of the most prominent partitioning clustering algorithms, it is selected to separate the projected feature-value pairs into clusters and obtain the corresponding medoid for each cluster. The whole pre-processing scheme is depicted in Fig 1, FC represents a feature cluster that is converted from feature-value pair cluster based on majority vote. In this case the number of feature cluster is 4. Consequently, four classification models are built on the resulted feature clusters. The threshold can be chosen from MAP values evaluated from classification results. Mean Average Precision (MAP) is the mean of average precision (AP) of all concepts and has been shown to have especially good discrimination and stability among evaluation methods. If using the selected threshold to eliminate the redundant classification models will produce the highest MAP in detecting semantic concept, then the same threshold will be used to remove unproductive classification models in testing section. Adjustment parameter π, classification model reliability α and the correlation of an interval of scores from each classification model to the target concept β were already proposed in [7]. In this paper, we introduce a new variable γ, which represents the correlation between cluster's medoid and the target concept, to be combined with the existing variable α using weighting factors. The optimal weighting factors can be obtained through logistic regression within the training data

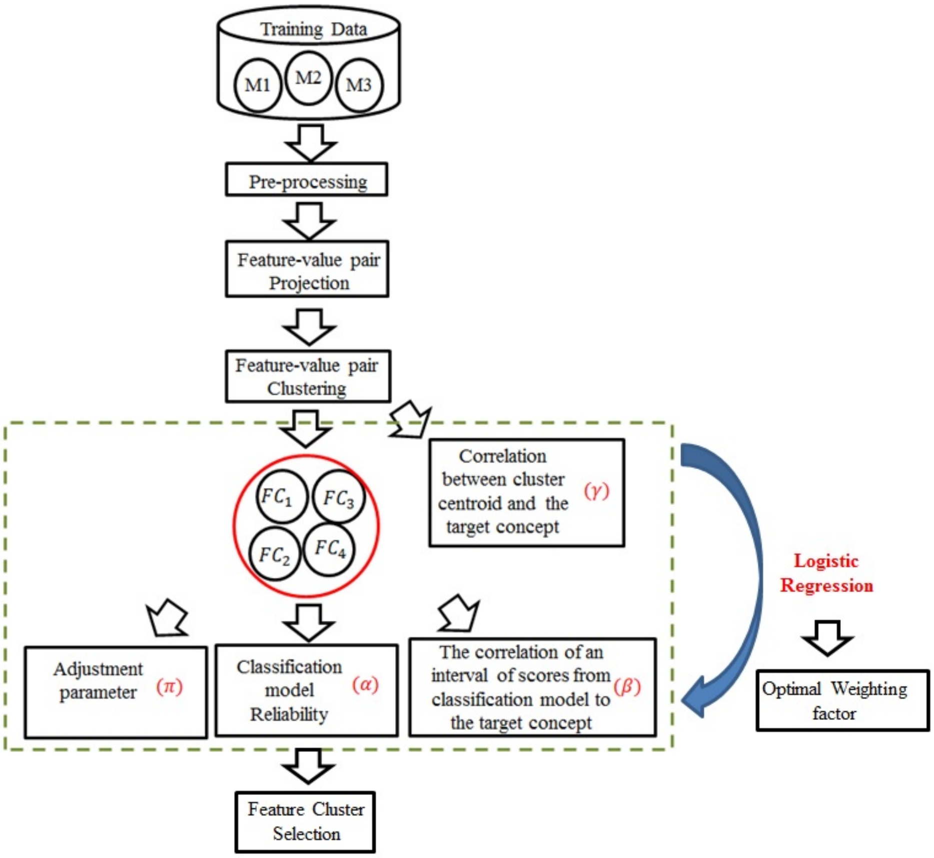

Training Data
M1
M2
M3
Pre-processing
Feature-value pair Projection
Feature-value pair Clustering
Correlation between cluster centroid and the target concept
(γ)
FC_1
FC_3
FC_2
FC_4
Logistic Regression
Optimal Weighting factor
Adjustment parameter
(π)
Classification model Reliability
(α)
The correlation of an interval of scores from classification model to the target concept
(β)
Feature Cluster Selection

Figure 1: Training section of MMC framework

set. The deriviation of variable γ will be covered in Section III-D.

B. Testing section of MMC framework

In the testing section, features extracted from multiple modalities are also converted into the same feature clusters as described in the training section. Unuseful feature clusters are eliminated according to the threshold set up in the training section. Four variables and two weighting factors attained from training data set are utilized in the proposed cluster-based ARC fusion method to improve model fusion and semantic concept detection.

C. Feature cluster selection

As mentioned in section III-A, after performing feature extraction on all the involving modalities followed by pre-processing step, MCA is applied to analyze the correlation among feature-value pairs by projecting them onto two major principal components as shown in Fig. 2. Each blue point represents a feature-value pair and the green square points are the medoids of feature clusters obtained after feature-value pair clustering. The positive class and the negative class are represented by a red triangle and a yellow circle respectively. Once the feature-value pair clusters are converted into feature clusters based on majority vote, one classification model will is built for each feature cluster. The ranking scores produced from the classification models are evaluated in terms of MAP value, which is considered as classification model reliability meaning how reliable each classification model it is to accurately detect the target concept. It is also the criterion to set up the threshold in eliminating unuseful classification models. As illustrated in section III-A, the threshold for each concept is decided based on the training section results. In addition, according to the observation of experimental results, feature clusters contain less than five features are also removed because this kind of classification models are lack of distinguishing capability.

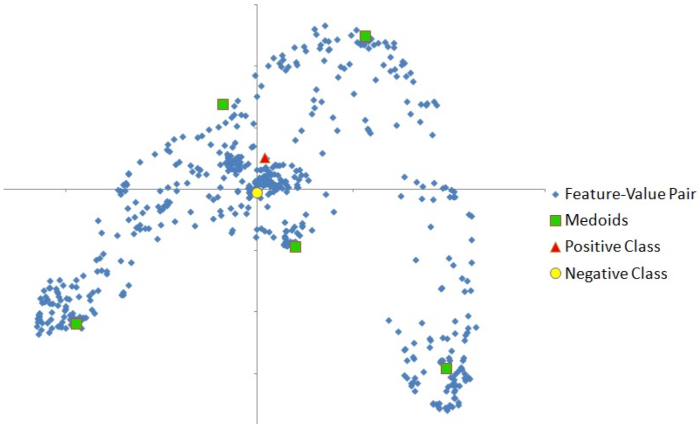


Figure 2: Feature-value pair projection and K-Medoid clustering results on a symmetric map

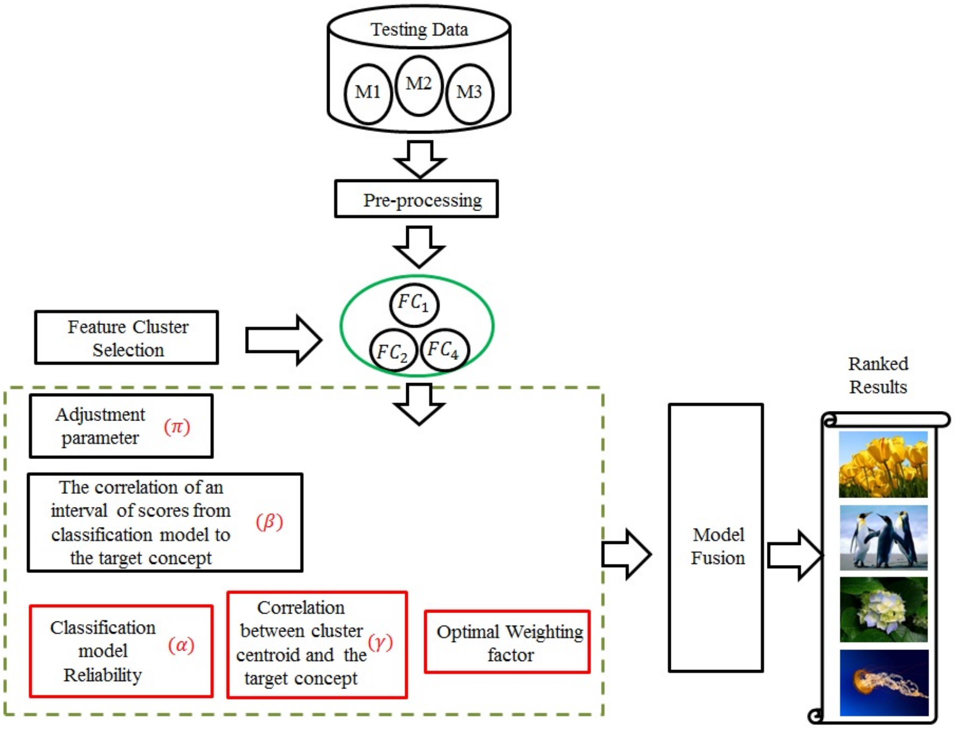

Testing Data
M1
M2
M3
Pre-processing
Feature Cluster Selection
FC_1
FC_2
FC_4
Adjustment parameter (π)
The correlation of an interval of scores from classification model to the target concept (β)
Classification model Reliability (α)
Correlation between cluster centroid and the target concept (γ)
Optimal Weighting factor
Model Fusion
Ranked Results

Figure 3: Testing section of MMC framework

D. Cluster-based ARC

The proposed cluster-based ARC is extended by introducing one new factor to better explore the correlation between models and concepts. The process will be depicted in two steps: section III-D1 describes how to generate variable γ and section III-D2 clarifies how to integrate γ into model fusion.

1) How to generate variable γ: As shown in Fig. 4, the symmetric map is the graphical representation after applying MCA and it can be used to visualize the medoids of feature clusters, positive class and negative class as points in a map with two dimensions, which are the first two principal components. The correlation between a medoid and a concept can be measured by the cosine value of the angle between the two vectors representing medoid and the positive class of the target concept. For example, the medoid of the third feature cluster is represented as CM_1 in Fig. 2, where Pos is the positive class and Neg is the negative class. θ_3 is the angle between CM_3 and the positive class Pos. If the absolute value of the $\cos(\theta_3)$ is large then it indicates a high correlation between the medoid CM_3 and the positive class.

As shown in equation 1, the cosine value can be obtained by using the inner product of the two vectors, e.g. medoid and positive class, and then it will be divided by the product of two vectors' length. The value of $\cos\theta_3$ is assigned to γ_i, which will be later integrated into fusion process.

$$\gamma_i = \cos(\theta)_i = \frac{Pos \cdot M_i}{\| Pos \| \| M_i \|} \tag{1}$$

2) How to integrate variable γ into model fusion: Given γ and α, which represent the correlation between model and concept and model reliability respectively, a weighting function is proposed as equation 2 to combine the two factors since they both indicate the importance of each classification model MI.

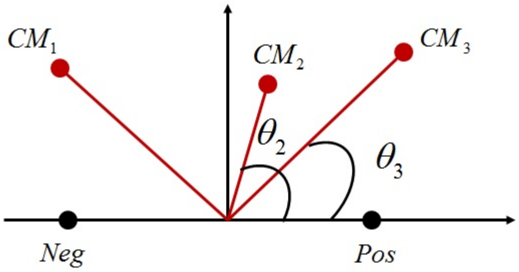


Figure 4: The symmetric map of the first two principal components

$$MI = \alpha \times \Lambda + \gamma \times (1 - \Lambda) \tag{2}$$

To combine the two factors with the most suitable weighting factor Λ, logistic regression is applied to produce the optimal weighting factor, which has been proved to minimize the error of semantic detection for each concept. Please notice that both α and γ are normalized through z-score normalization method before equation 2.

$$R(X|C) = \sum_{m=1}^{M} \frac{R_m(X|C)}{\pi} \left(\frac{MI_m \cdot \beta_m}{MI_m + \beta_m}\right) \tag{3}$$

Finally, given variable MI generated from equation 2, it replaced model reliability to have a harmonic balance with β, and normalization process was applied by introducing π, where it is the mean ranking score from different models to balance the score for X instance. $R(X|C)$ represents the final ranking score for instance X in detecting concept C. In equation 3, M is the number of classification models after feature cluster selection and β represents the correlation of an interval of scores from a ranking model to the target concept as described in [7].

In the next section, several experiments are conducted to compare the proposed fusion method with other fusion methods, i.e. min, max, mean, average, median, and ARC. The proposed cluster-based ARC was able to show that it better fuses the models by fully exploring the correlation among them and identifying model's importance in detecting concept.

IV. Experimental Analysis

A. Experiment Setup

To systematically evaluate the proposed framework, two experiments are designed: one demonstrates the fusion performance of the proposed framework, and the other one shows the improvement against our previous work. 3-fold cross validation and mean average precision (MAP) is applied on both experiments to validate the results of semantic concept detection. MCA is used as the classification modeling to evaluate the feasibility of the proposed framework [22][23][24].

In the first set of experiments, the performance of our proposed cluster-based ARC fusion method in semantic concept detection against several well-known fusion methods is demonstrated on both NUS-WIDE-LITE and NUS-WIDE 270K datasets [25]. NUS-WIDE data set is one of the largest real-world web image datasets including over 269,000 images with the ground-truth information of 81 concepts. With regard to visual features, most common visual features such as color histogram, edge direction histogram, wavelet texture, and bag of words based on SIFT descriptions are included with the dataset. In addition, all the images have their associated tags from flickr to represent as textual features. Therefore, it is the perfect benchmark dataset for multimedia semantic concept detection.

In the second set of experiments, the proposed MMC framework is also compared against our previous works, e.g. (CFA-MMF)[26] and (FCC-MMF)[6], and the original flat concatenation of multi-modality features, which is exact the same feature set but it is only trained as one classifier and there is no fusion process involved, on NUS-WIDE-LITE dataset.

B. Evaluation of Cluster-based ARC fusion method

Five reputable fusion methods including, minimum (min), maximum (max), average, mean, and median are adopted to be compared with the proposed cluster-based fusion method. The comparative experimental result demonstrated on NUS-WIDE LITE dataset is shown on Fig. 5a. It is observed that cluster-based ARC outperforms other well-known fusion approaches up to 25%. The positive difference between cluster-based ARC and the original ARC indicates the improvement in detecting semantic concepts. The lowest MAP value produced by median method is only 11% and our proposed method outperforms by 25%.

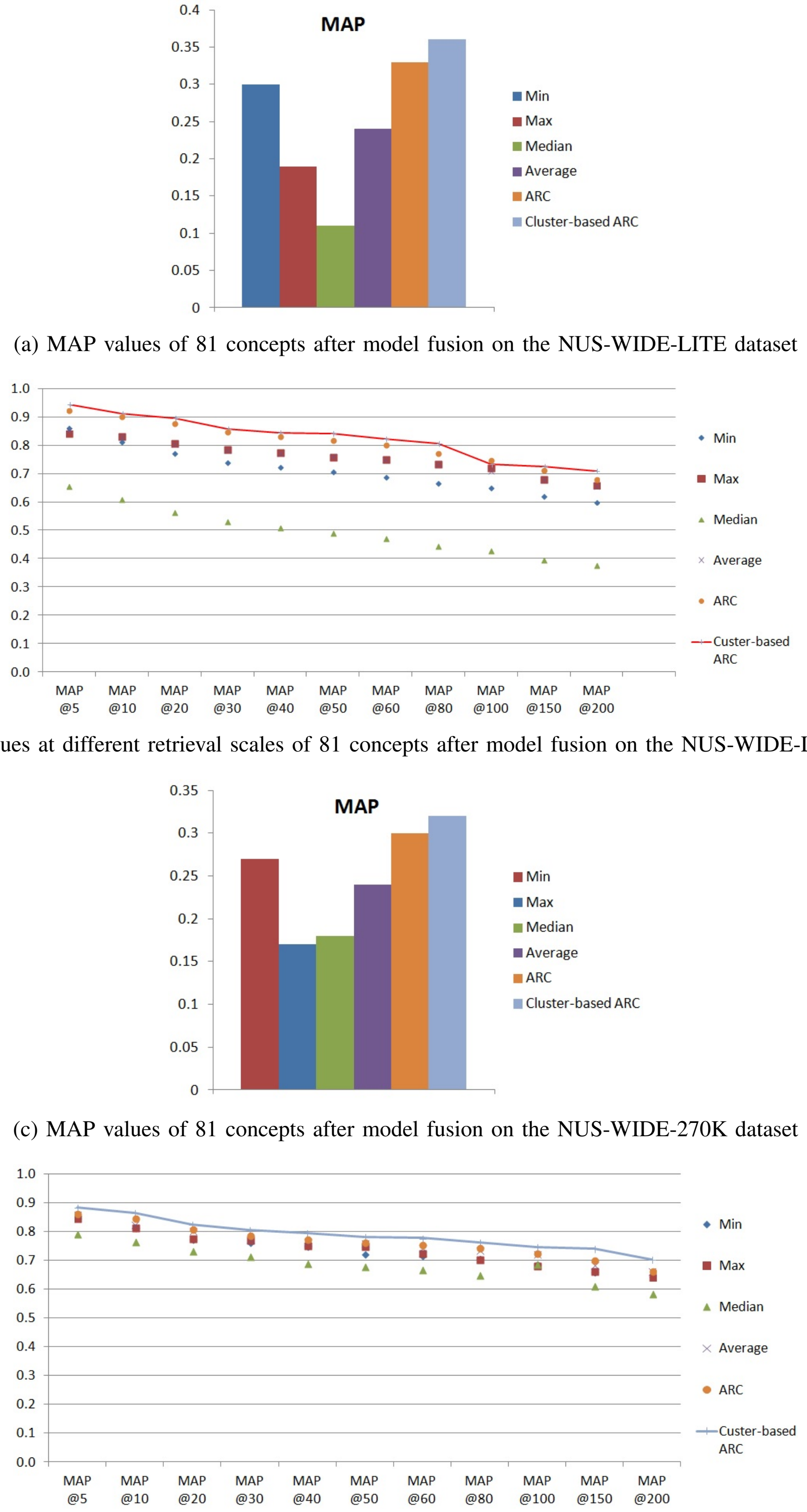
Fig. 5c shows the comparison results of the above-mentioned fusion methods on NUS-WIDE-270K dataset. The comparative results are quite similar as shown in Fig. 5a, however the relatively large size of the dataset has resulted in lower MAP values for all the fusion methods. Our proposed method was again validated to obtain the highest MAP value, where it is 15% and 2% higher than the worst and the best performance respectively. In addition, the performance of cluster-based ARC still outperforms the original ARC.

C. Evaluation of MMC framework

The proposed MMC framework is compared against our previous works, e.g. Feature Correlation Clustering-based Multi-Modality Fusion Framework (FCC-MMF) and Correlation-Based Feature Analysis and Multi-Modality Fusion Framework (CFA-MMF), and the original feature set, which simply combines all the features into one classification model, to validate whether our research work has been continuously advanced to adequately detect semantic concepts from multimedia data. The framework comparisons are carried out on NUS-WIDE-LITE dataset. As shown in Fig. 6 and Fig. 7, the proposed framework outperforms the previous works up to 20% in the first three retrieval scales. With regard to this experiment, couple observations are listed as follows: CFA-MMF was enhanced by reducing feature dimension while remaining comparative performance against original feature set; FCC-MMF converted features from multiple modalities into high intra-correlation and low inter-correlation feature clusters and they were consecutively trained as classification models where fusion process was applied to produce the final ranking scores; the proposed framework MMC went beyond FCC-MMF and further considered how feature cluster's medoid correlated to semantic concepts to improve fusion process.

V. Conclusion

The paper presents a multi-model collaboration framework including an enhanced fusion method called cluster-based ARC to effectively detect semantic concepts from multimedia data. Because of the experience learnt from our previous works, exploring correlation among multi-modalities has been proven effective in multimedia semantic concept detection. Therefore, the association between classification models and semantic concepts is combined with model reliability to represents as model importance, which indicates how useful the model it is in detecting this semantic concept. The idea of using only the useful feature clusters is also introduced in our framework. The proposed framework aims at thoroughly exploiting all the possible correlation from multiple modalities to build up a multi-model collaboration in semantic concept detection. The experiments are conducted on NUS-WIDE-LITE and NUS-WIDE-270K datasets to evaluate the propose framework. The comparative experimental results of 3-fold cross validation showed that the proposed framework outperforms several well-known fusion methods and our previous works in terms of MAP values.



(a) MAP values of 81 concepts after model fusion on the NUS-WIDE-LITE dataset

(b) MAP values at different retrieval scales of 81 concepts after model fusion on the NUS-WIDE-LITE dataset

(c) MAP values of 81 concepts after model fusion on the NUS-WIDE-270K dataset

(d) MAP values at different retrieval scales of 81 concepts after model fusion on the NUS-WIDE-270K

Figure 5: MAP values after model fusion on NUS-WIDE dataset

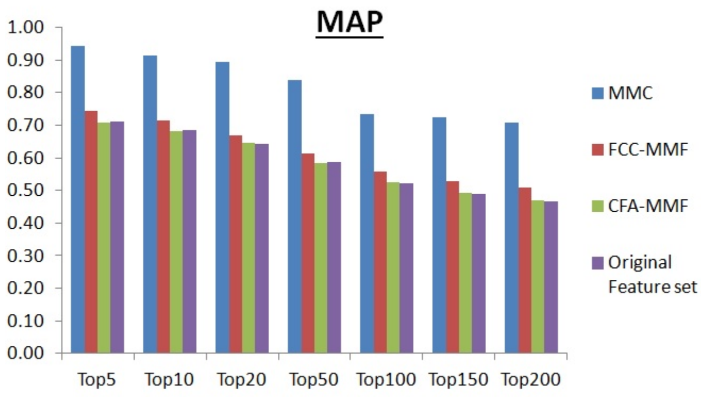
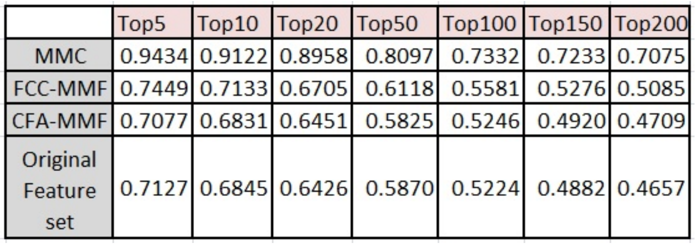


Figure 6: MAP values at different retrieval scales of 81 concepts of different frameworks on the NUS-WIDE-LITE dataset



	Top5	Top10	Top20	Top50	Top100	Top150	Top200
MMC	0.9434	0.9122	0.8958	0.8097	0.7332	0.7233	0.7075
FCC-MMF	0.7449	0.7133	0.6705	0.6118	0.5581	0.5276	0.5085
CFA-MMF	0.7077	0.6831	0.6451	0.5825	0.5246	0.4920	0.4709
Original Feature set	0.7127	0.6845	0.6426	0.5870	0.5224	0.4882	0.4657

Figure 7: MAP values at different retrieval scales of 81 concepts of different frameworks on the NUS-WIDE-LITE dataset

ACKNOWLEDGMENT

This research was supported in part by the U.S. Department of Homeland Security under grant Award Number 2010-ST-062-000039, the U.S. Department of Homeland Securitys VACCINE Center under Award Number 2009-ST-061-CI0001, and NSF HRD-0833093.

REFERENCES

[1] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[2] Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Dan Ellis, Shih-Fu Chang, Subhabrata Bhattacharya, and Mubarak Shah, "Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching.," in *TRECVID*, 2010.

[3] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 251–260.

[4] Hyun Oh Song, Stefan Zickler, Tim Althoff, Ross Girshick, Mario Fritz, Christopher Geyer, Pedro Felzenszwalb, and Trevor Darrell, "Sparselet models for efficient multiclass object detection," in *Computer Vision–ECCV 2012*, pp. 802–815. Springer, 2012.

[5] Gary Overett and Lars Petersson, "Large scale sign detection using hog feature variants," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*. IEEE, 2011, pp. 326–331.

[6] Hsin-Yu Ha, Fausto C Fleites, and Shu-Ching Chen, "Content-based multimedia retrieval using feature correlation clustering and fusion," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 4, no. 2, pp. 46–64, 2013.

[7] Chao Chen, Qiusha Zhu, Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen, "Web media semantic concept retrieval via tag removal and model fusion," *ACM Transactions on Intelligent Systems and Technology*, 2013.

[8] Henning Mèuller, Paul Clough, Thomas Deselaers, and Barbara Caputo, *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*, vol. 32, Springer, 2010.

[9] Alan F. Smeaton, Paul Over, and Wessel Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2006, pp. 321–330, ACM Press.

[10] William Hersh, Jayashree Kalpathy-Cramer, and Jeffery Jensen, "Medical image retrieval and automated annotation: Ohsu at imageclef 2006," in *Evaluation of Multilingual and Multi-modal Information Retrieval*, pp. 660–669. Springer, 2007.

[11] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[12] Abhishek Nagar, Karthik Nandakumar, and Anil K Jain, "Multibiometric cryptosystems based on feature-level fusion," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1, pp. 255–268, 2012.

[13] Jana Kludas, Eric Bruno, and Stephane Marchand-Maillet, "Information fusion in multimedia information retrieval," in *Adaptive Multimedial Retrieval: Retrieval, User, and Semantics*, pp. 147–159. Springer, 2008.

[14] Nan Luo, Zhenhua Guo, Gang Wu, and Changjiang Song, "Multispectral palmprint recognition by feature level fusion," in *Recent Advances in Computer Science and Information Engineering*, pp. 427–432. Springer, 2012.

[15] Mehmet Emre Sargin, Yücel Yemez, Engin Erzin, and A Murat Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[16] Ming Liu, Yun Fu, and Thomas S Huang, "An audio-visual fusion framework with joint dimensionality reducton," in *Acoustics, Speech*

and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. IEEE, 2008, pp. 4437–4440.

[17] Xiaona Xu and Zhichun Mu, "Feature fusion method based on kcca for ear and profile face based multimodal recognition," in *2007 IEEE International Conference on Automation and Logistics*. IEEE, 2007, pp. 620–623.

[18] Hervé Bredin and Gérard Chollet, "Audio-visual speech synchrony measure for talking-face identity verification," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 2, pp. II–233.

[19] Bakkama Srinath Reddy, "Evidential reasoning for multimodal fusion in human computer interaction," M.S. thesis, University of Waterloo, 2007.

[20] Hatice Gunes and Massimo Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *Systems, Man and Cybernetics, 2005 IEEE International Conference on*. IEEE, 2005, vol. 4, pp. 3437–3443.

[21] Gerasimos Potamianos, Chalapathy Neti, and Sabine Deligne, "Joint audio-visual speech processing for recognition and enhancement," in *AVSP 2003-International Conference on Audio-Visual Speech Processing*, 2003.

[22] Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen, "Correlation-based interestingness measure for video semantic concept detection," in *IEEE International Conference on Information Reuse & Integration, 2009. IRI'09.* IEEE, 2009, pp. 120–125.

[23] Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen, "Enhancing concept detection by pruning data with mca-based transaction weights," in *11th IEEE International Symposium on Multimedia, 2009. ISM'09.* IEEE, 2009, pp. 304–311.

[24] Lin Lin, Chao Chen, Mei-Ling Shyu, and Shu-Ching Chen, "Weighted subspace filtering and ranking algorithms for video concept retrieval," *MultiMedia, IEEE*, vol. 18, no. 3, pp. 32–43, 2011.

[25] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2009, p. 48.

[26] Hsin-Yu Ha, Yimin Yang, Fausto C Fleites, and Shu-Ching Chen, "Correlation-based feature analysis and multi-modality fusion framework for multimedia semantic retrieval," in *2013 IEEE International Conference on International Conference on Multimedia and Expo (ICME), "Multimedia for Humanity" Theme Track*, JUL 2013.