Multimedia Fusion With Mean-Covariance Analysis

Xiangyu Wang and Mohan S. Kankanhalli

***Abstract*—The number of multimedia applications has been increasing over the past two decades. Multimedia information fusion has therefore attracted significant attention with many techniques having been proposed. However, the uncertainty and correlation among different information sources have not been fully considered in the existing fusion methods. In general, the predictions of individual information source have uncertainty. Furthermore, many information sources in the multimedia systems are correlated with each other. In this paper, we propose a novel multimedia fusion method based on the portfolio theory. Portfolio theory is a widely used financial investment theory dealing with how to allocate funds across securities. The key idea is to maximize the performance of the allocated portfolio while minimize the risk in returns. We adapt this approach to multimedia fusion to derive optimal weights that can achieve good fusion results. The optimization is formulated as a quadratic programming problem. Experimental results with both simulation and real data confirm the theoretical insights and show promising results.**

***Index Terms*—Sensor fusion, portfolio theory, data analysis.**

I. Introduction

THE number of multimedia applications has been increasing over the past few decades. Consequently, multimedia data analysis has attracted significant attention. Multimedia data generally comprise of data of different information sources. A multimedia analysis task involves processing of multimodal data in order to obtain valuable insights about the data, a situation, or a higher level of activity [1]. For example, surveillance systems utilize the data from multiple types of sensors like microphones, video cameras, etc., to detect certain events. For news video retrieval, video data is combined with audio data and text information to enable content-based search. In most applications, no single information source can accomplish the analysis task perfectly. Hence we use multimedia fusion to integrate multiple modalities, their associated features, or the intermediate decisions in order to perform the task. One of the advantages of multimedia fusion is to use the correlation of different information sources. Thus, how to measure and combine the correlation appropriately is an important problem.

Generally speaking, there are three different fusion categories: data, feature, and decision level fusion, depending on the processing stage at which fusion takes place [2]. Unlike data and feature level fusion where the data or features are usually heterogeneous and hard to combine, the decisions are homogeneous in nature. The data from different modalities can be analyzed using different yet appropriate methods to obtain the unimodal decisions. This provides much more flexibility in the multimodal fusion process. Moreover, for decision level fusion, it is easy to control the relative contributions of information sources to fusion results (e.g., by weighting), while this is more difficult in data and feature level fusion. Thus, decision level fusion is more appropriate in the context of multimedia fusion, and we will focus on decision level fusion in this paper.

A. Multimedia Decision Fusion Strategies

There are generally two strategies widely used for combining decisions of multiple information sources: Linear Opinion Pool and Independent Opinion Pool. By decision, we mean the output of a classification model. Given an observation, the decision is how likely the observation belongs to a predefined class. In this paper, it is the probabilistic output of a classifier.

- Linear Opinion Pool (LOP) is proposed by Stone in [3]. By attaching a measure of value such as weight to the information provided by each source, the decisions from each information source are combined linearly [4]. Let $\mathbf{x}^{(i)}$ be the observations from the ith source, $I(\bullet)$ be the decision, y be the class, and n be the number of sources. It is defined as:

$$I\left(\mathbf{y}\left|\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(n)}\right.\right)\propto\sum_{i=1}^{n}w_i I\left(\mathbf{y}\left|\mathbf{x}^{(i)}\right.\right) \quad (1)$$

  where w_i is a weight such that, $0 \leq w_i \leq 1$ and $\sum_{i=1}^{n} w_i = 1$. The weight w_i reflects the significance attached to the ith source.
- Independent Opinion Pool (IOP) is derived in [5] by assuming the information obtained conditioned on the observation set $p(\mathbf{y} \mid \mathbf{x}^{(i)})$ is independent. It is defined as:

$$I\left(\mathbf{y}\mid\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(n)}\right)\propto\prod_{i=1}^{n}I\left(\mathbf{y}\mid\mathbf{x}^{(i)}\right) \quad (2)$$

  Here, $I(\mathbf{y} \mid \mathbf{x}^{(i)}) = p(\mathbf{y} \mid \mathbf{x}^{(i)})$ is the probabilistic decision for y obtained based on the observations from the ith source.

The work in [6] concluded that the IOP works well only when the posterior probability $p(\mathbf{y} \mid \mathbf{x}^{(i)})$ of individual classifiers can be accurately estimated. The LOP strategy is more tolerant to noise because the sum does not magnify noise as severely as the product [7]. If there are dependencies between information sources, the LOP should be used instead of the IOP [4]. Kittler *et al.* compared different classifier combination schemes in [8]. It is concluded that LOP is more restrictive in assumptions (it assumes that the a posteriori probabilities computed by the respective classifiers will not deviate dramatically from the prior prob-

Manuscript received December 10, 2010; revised September 24, 2011, March 06, 2012, and April 03, 2012; accepted May 22, 2012. Date of publication October 16, 2012; date of current version December 12, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel Gatica-Perez.

The authors are with National University of Singapore, Singapore.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2012.2225027

abilities) but more robust than the IOP. It is worth mentioning that LOP applies for a wider class of functions. For comparison, we have used the appropriate functions whose "decision" is a probability density in this paper.

Logarithmic Opinion Pool (LGP) [9] is also a popular fusion strategy. It interprets the decision as a probability statement, and defines the distance to be the Kullback-Leibler divergence. Then, linear averaging of the outputs corresponds to logarithmic averaging of the probability statements:

$$I\left(\mathbf{y} \,\middle|\, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\right) \propto \prod_{i=1}^{n} I\left(\mathbf{y} \,\middle|\, \mathbf{x}^{(i)}\right)^{w_i} \quad (3)$$

After making certain approximations, Heskes *et al.* [9] proposed to select the weights by minimizing:

$$K(q, I) \approx \sum_i w_i K(q, I) - \frac{1}{4} \sum_{i,j} w_i w_j [K(I_i, I_j) + K(I_j, I_i)] \quad (4)$$

Here, q is the true probability, K is the Kullback-Leibler divergence, and I_i is the decision based on the observations from the ith source. A drawback is that the pool assigns a probability of zero if any source assigns a zero value. For multimodal fusion, there may not be ways to measure the individual probabilities accurately. Moreover, the individual probabilities are estimated from the data of different information sources and not from the same data.

There are also many training-based fusion methods. Wu *et al.* proposed training-based super-kernel fusion in [7]. It first finds statistically independent modalities from the raw features. Then, their method determines the optimal combination of information sources by training the output scores of different information sources. It needs additional processing of the data and training for combinations of information sources, which is computationally expensive. Moreover, the method is not scalable. The models need to be trained again when a new information source is introduced.

Evidence theory has also been used in multimedia fusion such as in [10], [11]. It has been found to be more suitable for handling mutually inclusive hypotheses. However, the *mass functions* and *belief functions* need to be designed for different applications, which degrades the portability of evidence theory. It also suffers from the combinatorial explosion and conflicting beliefs problem.

As stated above, the linear fusion strategy is computationally less expensive compared to other strategies [1]. It is one of the simplest, most widely used methods and is also easily scalable. Several methods based on linear fusion have been proposed. Max/min/average fusion [12] takes the maximum/minimum/average prediction score of all information sources as the final prediction score. However, these methods do not take the difference between performances of information sources into account. Weighted fusion [13] obtains the final decision by assigning the logarithm of accuracy-error rate as weights to different information sources. In general, most of the linear fusion methods consider the fusion as an information aggregation task. They try to maximize the aggregated information by assigning proper weights to individual information sources [14]. However, finding the appropriate weights for different information sources is still an open research issue [1].

In addition, there are also some related works such as regression aggregation in statistics. Bunea *et al.* [15] studies statistical aggregation procedures in the regression setting. The convex aggregation is to select the optimal convex combination of the given estimators. Suppose $f_1, \ldots, f_n \in [-L, L]$ are the given estimators, the aggregate $\tilde{f}$ is defined as:

$$\tilde{f} = \sum_{i=1}^{n} w_i f_i \quad (5)$$

where $w_i \geq 0$, $\sum_{i=1}^{n} w_i \leq 1$. Juditsky *et al.* [16] proposed algorithms to find an aggregated estimator that is nearly as good as the best convex combination. Via the stochastic counterpart approach, the authors found the optimal weights by minimizing $\mathbf{w}^{\mathrm{T}} A \mathbf{w} - b^{\mathrm{T}} \mathbf{w}$. To apply it in multimedia fusion, we take $f_i = I_i$. Here,

$$A_{ij} = \frac{1}{T} \sum_{t=1}^{T} I_i(\mathbf{x}_t) I_j(\mathbf{x}_t) \quad (6)$$

$$b_i = 2 \frac{1}{T} \sum_{t=1}^{T} y(\mathbf{x}_t) I_i(\mathbf{x}_t) \quad (7)$$

Logistic regression is used for prediction of the probability of occurrence of an event by fitting data to a logistic function. By minimizing the error, the method obtains coefficients $\beta_i, i = 1, \ldots, n$ so that

$$I(\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i I_i(\mathbf{x}))}} \quad (8)$$

It can also be used to combine the decisions from different information sources linearly. It is obvious that the method only tries to minimize the error. It can be sensitive to the noisy information source.

Perfect classification models generally cannot be learned due to noise and the training/test distribution gap. Moreover, in a multimedia data-understanding task, we often assert similarity between data based on our beliefs which does not come from classical probability experiments [7]. There are many sources of uncertainty such as ambiguity, noise, and deviations between the scoring function and the true probability of relevance. Thus, the risk (uncertainty) is an intrinsic feature prediction using classification models. Taking the real accuracy as type of "an investment return" of our classification models, we should maximize the return and minimize the variance of the return. To the best of our knowledge, minimizing the effect of uncertainties has never been explicitly considered in multimedia fusion methods. To solve this problem, the portfolio theory is employed. A preliminary version of portfolio fusion has been described in [17]. Simulation and concept detection experiments have shown that it outperforms average fusion, weighted fusion, and naive Bayesian fusion method. Here, more description and details with varying return definitions are shown. Furthermore, experiments are done to show the superiority of proposed portfolio fusion method to other related fusion methods.

In the rest of this paper, we will first introduce the portfolio theory and its applications in Section II. Then, we formulate and describe the multimedia portfolio fusion method in Sections III and IV. Sections V, VI and VII show the experimental results. The proposed portfolio fusion is based on linear fusion, and it is compared with several related linear fusion methods to demonstrate the superiority. Finally, the discussion and conclusions are given in Section VIII.

II. Portfolio Theory

Portfolio theory was introduced by Markowitz in [18], for which he won the Nobel prize in economics. Portfolio theory is a theory of investment which tries to maximize the return and minimize the risk by carefully choosing different securities and is widely used in the finance industry. Security is a legal entitlement to receive an amount of money. A portfolio is a combination of existing securities, which tells us how many units of each security have to be bought or sold to create the portfolio [19]. The portfolio theory models a portfolio as a weighted combination of securities. It starts with relevant beliefs about future performances of available assets according to the past track records, and ends with a choice of portfolio. There are two salient features of any security investment.

- Uncertainty is an inherent feature of security investment. Economic forces are not understood well enough for predictions to be beyond doubt or error. Moreover, non-economic influences can change the success of a particular investment.
- The correlation among security returns is another inherent feature of security investments. To reduce investment risk, it is necessary to avoid a portfolio whose securities are all highly correlated with each other. One hundred securities whose returns rise and fall in near unison afford little more protection than the uncertain return of a single security.

The criteria for choosing an investment portfolio, which serves as a guide to the important and unimportant, the relevant and irrelevant, depends on the nature of the investor. Investors can be conservative, balanced, or aggressive based on their appetite for risk. Conservative investors may emphasize more on low risk. However, two objectives of portfolio analysis are common:

- Investors want the "return" to be high.
- They want this return to be dependable, stable, not subject to uncertainty.

By combining different securities whose returns are not correlated, portfolio theory seeks to reduce the total variance of the portfolio. The appetite for risk determines the variance. Due to its excellent performance, this theory is widely used today.

The multimedia fusion problem is quite similar to portfolio analysis in finance. Each information source can be considered as a security in financial investment. The information source also has two salient features.

- First of all, the information source has the uncertainty feature. There is still no perfect learning to predict without doubt or error.
- Moreover, it is quite common that some information sources are correlated.

In multimedia fusion, we study each information source to obtain classifiers and invest different weights on each information source to obtain good classification results on future unseen instances. The objective is to achieve a high dependable return. Most of existing methods aim to maximize the expected accuracy. However, it will not be able to guarantee real high accuracy due to uncertainty. Even when we have a classifier with high expected accuracy, it is not safe if its variance is high [20]. Thus, uncertainty is an extremely important feature that demands serious consideration. Moreover, the information sources in the multimedia systems are generally correlated. It is not always correct to assume independence of the modalities. Thus, diversification is beneficial for multimedia fusion. Poh *et al.* in [21] discussed how the correlations affect the fusion performance. It is shown that the more dependent the information sources are, the lesser the gain one can benefit out of fusion. The positive correlation "hurts" fusion (fusing two correlated information sources of similar performance will not always beneficial) while negative correlation (greater "diversity") improves fusion. As stated above, both uncertainty and correlation should be considered in multimedia fusion. It is not advisable to only maximize the expected performance in multimedia fusion. We should attempt to maximize expected accuracy and minimize risk to achieve an overall good performance. The portfolio theory is key in helping achieve this.

III. Problem Formulation

- S is a multimedia system designed for accomplishing a task D. The multimedia system S consists of $n \geq 1$ correlated information sources $M_1, M_2, \ldots, M_n$.
- Let $y(\mathbf{x})$ be the actual class of instance $\mathbf{x}$. In this paper, $y(\mathbf{x}) : \mathbf{x} \rightarrow \{-1, 1\}$ for the tasks appeared in this paper: simulation classification, concept detection and human detection.
- For $1 \leq i \leq n$, let $I_i(\mathbf{x})$ be the prediction of the task D based on information source M_i on instance $\mathbf{x}$. It is obtained by employing a classifier on the features extracted from M_i, and can be either probabilistic output (posterior probability estimates) or decision output (belief values or $-1/+1$ decision values). The final prediction I of S consisting of information sources $M_1, M_2, \ldots, M_n$ is modeled as:

$$I(\mathbf{x}) = \sum_{i=1}^{n} w_i I_i(\mathbf{x}) \tag{9}$$

  where w_i is the normalized weight assigned to M_i. $0 \leq w_i \leq 1, \sum_{i=1}^{n} w_i = 1$. In the binary classification problem, the prediction $I_i(\mathbf{x})$ is the output of likelihood for the target hypothesis from the classifier trained based on the observations in M_i. $I_i(\mathbf{x}) : \mathbf{x} \rightarrow [0, 1]$, and $I_i(\mathbf{x}) \geq 0.5$ indicates presence of the target hypothesis.
- For $1 \leq i \leq n$, let $r_i(\mathbf{x})$ be the return of M_i at $\mathbf{x}$ individually, and R_i be the expected return of M_i, which is defined as $R_i = E[r_i]$. The return depends on the application.
- For $1 \leq i, j \leq n$, let $\Phi = [\Phi_{ij}]$ be the covariance matrix between information sources. The element Φ_{ij} is defined as $\Phi_{ij} = E[(r_i - E[r_i])(r_j - E[r_j])]$. It captures the correlations of different information sources.

Our aim is to find the optimized weights w_i so that the fusion prediction I achieves good performance (desirable results, e.g., high accuracy or high average precision, for different applications). In this paper, we focus on the classification and retrieval applications. For the classification application, we aim to achieve high accuracy. For the retrieval application, we aim to achieve high mean average precision.

To solve this problem, the portfolio theory is employed to obtain suitable weights. The portfolio theory helps pick a portfolio of securities according to their return and risk. To adopt it in multimedia fusion, the return and risk of each information source need to be defined first.

IV. Proposed Approach

A. Return and Risk

Each information source in the multimedia system is considered the equivalent of a security in financial investment. The definition can be varied to different applications according to their aims. For example, in the retrieval problem, where we evaluate the performance using average precision, the definition of return can be:

$$r_i(\mathbf{x}) = (I_i(\mathbf{x}) - 0.5)y(\mathbf{x}) \tag{10}$$

The return of portfolio is $r = \sum_{i=1}^{n} w_i r_i = (\sum_{i=1}^{n} w_i I_i - 0.5)y = (I - 0.5)y$. Better return means better decision. For $y = 1$, better return means larger decision value I. For $y = -1$, better return means smaller decision value I. Thus, finding the optimal weights with return results in the optimal weights for combining decisions.

In the classification problem, since the aim of the classifier of the information source is to accurately predict the labels and the performance is evaluated using accuracy, the return should be positive if the prediction is correct and negative otherwise. For M_i on instance $\mathbf{x}$, the return $r_i(\mathbf{x})$ is defined as:

$$r_i(\mathbf{x}) = h_i(\mathbf{x})y(\mathbf{x}) \tag{11}$$

where $h_i(\mathbf{x}) = \text{sign}(I_i(\mathbf{x}) - 0.5) = (I_i(\mathbf{x}) - 0.5)/(|I_i(\mathbf{x}) - 0.5|)$ is the predicted class of M_i on instance $\mathbf{x}$. Suppose that $\lambda_1 = \min\{|I_i - 0.5|, i = 1, \ldots, n\}$ and $\lambda_2 = \max\{|I_i - 0.5|, i = 1, \ldots, n\}$. Then, the return of portfolio is $r = \sum_{i=1}^{n} w_i r_i = ((I - 0.5)y)/(\lambda)$, where $\lambda_1 \leq \lambda \leq \lambda_2$. For $y = 1$, positive r means larger I; for $y = -1$, positive r means smaller I. Thus, finding the optimal weights with return results in the optimal weights for combining decisions to obtain correct prediction.

Based on the return definition, the expected return of ith information source R_i is approximated using $r_i(\mathbf{x})$ over all the previous instances $\mathbf{x}_t, t = 1, \ldots, T$:

$$R_i = E[r_i] = \frac{1}{T}\sum_{t=1}^{T} r_i(\mathbf{x}_t) \tag{12}$$

The risk of information source is modeled as the standard deviation σ of return. For M_i,

$$\sigma_i^2 = E[(r_i - E[r_i])^2] \tag{13}$$

$\sigma_i \in [0, 1]$, and a larger value indicates more risk.

B. Correlation

The correlation among different information sources represents how they co-vary with each other. Moreover, diversification which is related to correlation among information sources is beneficial for multimedia fusion to reduce risk.

The popular Pearson's correlation coefficient is used to measure the correlation between different information sources. Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. For M_i and M_j, the correlation ρ_{ij} is defined as:

$$\rho_{ij} = \frac{\Phi_{ij}}{\sigma_i \sigma_j} \tag{14}$$

$\rho_{ij} \in [-1, 1]$. The σ_i and σ_j are the standard deviation of returns for M_i and M_j, which is defined in (13). $\Phi = [\Phi_{ij}]_{n \times n}$ is the covariance matrix of n information sources, which captures the correlations and risk of multiple information sources.

C. Optimal Weights With Portfolio Theory

Using the portfolio theory, the optimal portfolio can be found by minimizing the following expression, which is to maximize the return while minimizing the variance of return:

$$f = \mathbf{w}^{\mathrm{T}}\Phi\mathbf{w} - \alpha\mathbf{R}^{\mathrm{T}}\mathbf{w} \tag{15}$$

where

- $\mathbf{w} = [w_1 \ldots w_n]^{\mathrm{T}}$ is the vector of the weights for information sources. $0 \leq w_i \leq 1$ and $\sum_{i=1}^{n} w_i = 1$.
- $\mathbf{R} = [R_1 \ldots R_n]^{\mathrm{T}}$ is the vector of the expected returns for information sources. $\mathbf{R}^{\mathrm{T}}\mathbf{w}$ captures the expected return of the portfolio of information sources.
- Φ is the covariance matrix for the information sources in the multimedia system. $\mathbf{w}^{\mathrm{T}}\Phi\mathbf{w}$ captures the variance of return of the portfolio.
- $\alpha \in [0, +\infty)$ is "risk tolerance" factor. Different values can be used for different "risk appetite" in various applications. $\alpha = 0$ results in minimizing the risk (conservative risk appetite), while $\alpha = +\infty$ results in maximizing the expected return of the fusion results (aggressive risk appetite).

To solve the optimization problem, we used the *quadratic programming* (QP) approach. In multimedia portfolio fusion method, the aim is to minimize

$$f = \mathbf{w}^{\mathrm{T}}\Phi\mathbf{w} - \alpha\mathbf{R}^{\mathrm{T}}\mathbf{w}$$

It is equivalent to minimizing (by multiplying $(1/2)$ on both sides)

$$f_{\mathbf{w}} = \frac{1}{2}f = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\Phi\mathbf{w} + \mathbf{R}_p^{\mathrm{T}}\mathbf{w}$$

where, $\mathbf{R}_p = (-(\alpha/2))\mathbf{R}$. There is one equality constraint ($\sum_{i=1}^{n} w_i = 1$) and n inequality constraints ($w_i \geq 0$ for $i = 1, \ldots, n$). In our case, Φ is the covariance matrix according to our definition. It can be proved that the covariance matrix is positive semidefinite. The problem is thus a convex QP and can be solved using the active set method.

Input: Labeled observations for n information sources, "risk tolerance" factor α
Output: Optimal weights $\mathbf{w}$ for the fusion
for all Information source M_i **do** // Calculate the return for each information source
Train a classification or decision model $model_i$ using the observations in information source M_i
Calculate the return $r_i(\mathbf{x})$ for each observation in information source M_i
Obtain the expected return R_i for information source M_i
end for
Obtain the vector of the returns for all the information sources $\mathbf{R} = [R_1 \cdots R_n]^{\mathrm{T}}$ // Calculate the return vector
for all Information source pair M_i, M_j **do** // Calculate the covariance matrix of information sources
Calculate the covariance Φ_{ij} of information source M_i, M_j
end for
Obtain the covariance matrix $\Phi = \left[\Phi_{ij}\right]_{n\times n}$
$\mathbf{R}_p = (-\frac{\alpha}{2})\mathbf{R}$
Minimize $f_{\mathbf{w}} = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\Phi\mathbf{w} + \mathbf{R}_p^{\mathrm{T}}\mathbf{w}$ using the constraints to obtain optimal weights $\mathbf{w}$ (*e.g.*, using Active Set Method)

Fig. 1. Optimal weights determination by portfolio theory.

Input: Optimal weights $\mathbf{w}$ for different information sources
Output: Prediction result c
In the future, for instance $\mathbf{x}$
Calculate the prediction $I(\mathbf{x}) = \sum_{i=1}^{n} w_i I_i(\mathbf{x})$ for each class
The class c with largest prediction value is the predicted result for $\mathbf{x}$

Fig. 2. Portfolio fusion method.

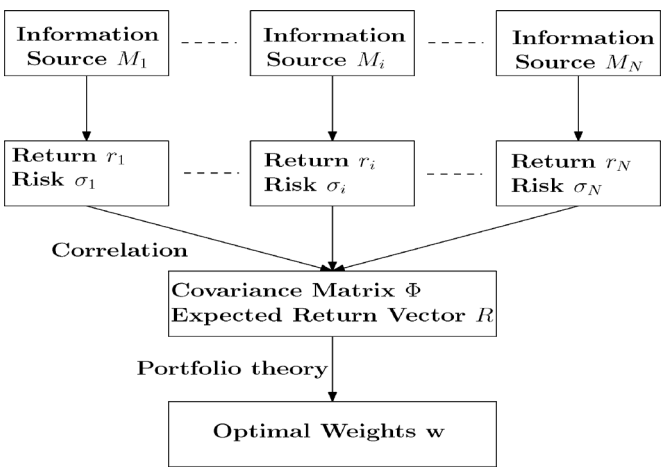


Fig. 3. Architecture of Portfolio Fusion Method.

D. Multimedia Portfolio Fusion

In summary, the portfolio fusion method is described in Figs. 1 and 2, and is illustrated in Fig. 3. The return and risk for information sources are first computed using the past (training) data. Then, portfolio theory gives the optimal weights for the information sources by minimizing the risk while maximizing the return, as shown in Fig. 1. After that, as shown in Fig. 2, for each test instance, the predictions from the information sources will be combined linearly using the optimal weights, and the class with largest prediction value will be chosen as the class of the instance. When a new information source is introduced, only the correlations will be computed against training the fusion model again in training-based fusion method, which saves many computation efforts.

Input: μ_0, μ_1, σ for n information sources
Output: Fusion results for $\{M\}_1^K, K = 1, ..., n$
for Repetition $l = 1$ to L **do**
for all Information source M_i **do** // Data generation and model training
Generate SN negative samples and SP positive samples
Select 50% negative and positive samples as training dataset, and use remaining negative and positive samples as test dataset
Train a classification model $model_i$ for information source M_i using the training data
end for
for all Combination of information sources $\{M\}_1^K$ **do** // Multimedia fusion
Find the optimal weights $\mathbf{w}$ using portfolio theory
Calculate classification performance and report the results
end for
end for
Report the average achieved performance over L repetitions

Fig. 4. The Simulation Procedure.

V. Simulation Experiments

To show the effectiveness of proposed portfolio fusion method (PTF), we simulated multiple information sources and compared with different fusion methods: logarithmic opinion pool (LGP), convex aggregation (CA), logistic regression (LR), and super kernel fusion method (SKF). The classification problem is considered here. The simulation was performed on the fusion of n different information sources. The notations are:

- Let $\{M\}_1^K, K = 1, 2, \ldots, n$ represent the fusion results of information sources 1 to K.
- Let $\mathcal{N}(\mu, \sigma)$ represent a Gaussian distribution with mean μ and standard deviation σ.

In general, the classification models are trained using SVM with LIBSVM [22]. The calibrated probabilistic outputs from SVMs using Platt Scaling [23], which is in (0, 1), is used. The function I_i in this case is the probabilistic output from SVM using Platt Scaling based on information source M_i, while the function h_i is the predicted label. The results of the simulation runs are described in the following. Section V.A gives the description of the simulation setup. Section V.B shows the performance with different configurations of information sources. Section V.C shows the performance with different values of risk tolerance factor.

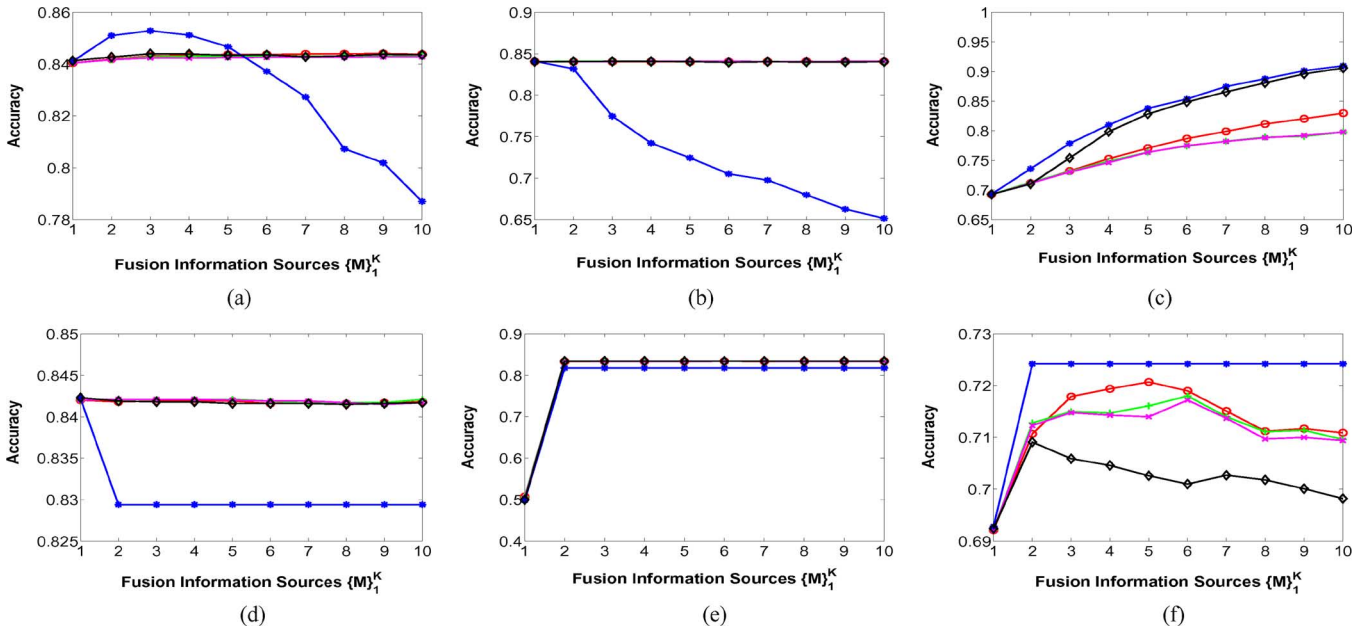


Fig. 5. Results of simulation runs for different simulation scenarios. Red circle represents PTF, green plus sign represents LGP, magenta cross represents CA, blue asterisk represents LR, black diamond represents SKF. (a) Results of scenario 1, (b) Results of scenario 2, (c) Results of scenario 3, (d) Results of scenario 4, (e) Results of scenario 5, (f) Results of scenario 6.

A. Simulation Setup

We simulate n information sources and then fuse them one by one. For each information source, we generally assume two Gaussians for the negative and positive class [24]. We randomly generate the annotated collection for each information source (which carries $-1/+1$ labels for each instance). The negative class has distribution $\mathcal{N}(\mu_0, \sigma)$ and positive class has distribution $\mathcal{N}(\mu_1, \sigma)$. The number of instances is $SP = SN = 100$. To reduce the effects of randomness in the results, we repeat every simulation run $L = 50$ times. The results of each simulation run is actually obtained as an average over 50 times. The actual simulation process is described in Fig. 4.

B. Simulation Parameter Variation

We choose different values for simulation parameters μ_0, μ_1, σ for each information source. In this way, we validate the performance of the proposed fusion algorithm in different scenarios. Here, a large standard deviation σ represents more noise thus more uncertainty in the data. The large mean difference between positive and negative class $\mu_1 - \mu_0$ represents more discrimination in the data. For each simulation run, we use new seeds for the random number generator to ensure high quality of randomness. Generally, the "risk tolerance" factor α is set such as the values in the covariance matrix Φ approximate the standard deviation to trade-off between risk and return. Here, we simply set $\alpha = 2$, which corresponds to a moderate appetite for risk of a balanced investor.

In the simulation experiments, we test the methods on 6 different scenarios with up to $n = 10$ information sources. The simulations are tested on both independent and correlated information sources. Then, in each of these cases, we examined the information sources with different standard deviation and same standard deviation (both large discriminative and small discriminative information sources). The six scenarios are specified as follows:

TABLE I
DESCRIPTIONS OF SIMULATION SCENARIOS: $\mu_1 - \mu_0$ DENOTES MEAN DIFFERENCE BETWEEN POSITIVE AND NEGATIVE CLASS, σ DENOTES STANDARD DEVIATION

Scenario	Correlation	$\mu_1 - \mu_0$	σ
1	Independent	2	i for M_i
2	Independent	2	1 for M_1, 16 for M_2–M_{10}
3	Independent	1	1
4	Correlated (M_2–M_{10})	2	1 for M_1, 16 for M_2–M_{10}
5	Correlated (M_2–M_{10})	2	16 for M_1, 1 for M_2–M_{10}
6	Correlated (M_2–M_{10})	1	1

1) Independent (different standard deviation for all information sources)
2) Independent (different standard deviation for some information sources)
3) Independent (same standard deviation for each information source)
4) Correlated (different standard deviation for two categories of information sources, less noisy information sources)
5) Correlated (different standard deviation for two categories of information sources, more noisy information sources)
6) Correlated (same standard deviation for two categories of information sources)

The descriptions can be found in Table I. The simulation results are shown in Fig. 5.

We can draw the following conclusions from the simulation results:

- For independent information sources (scenario 1–3),
  - For the information sources that are of the similar performance (scenario 3), the fusion results of portfolio fusion outperforms logarithmic opinion pool and convex aggregation method, while it is a little worse than the

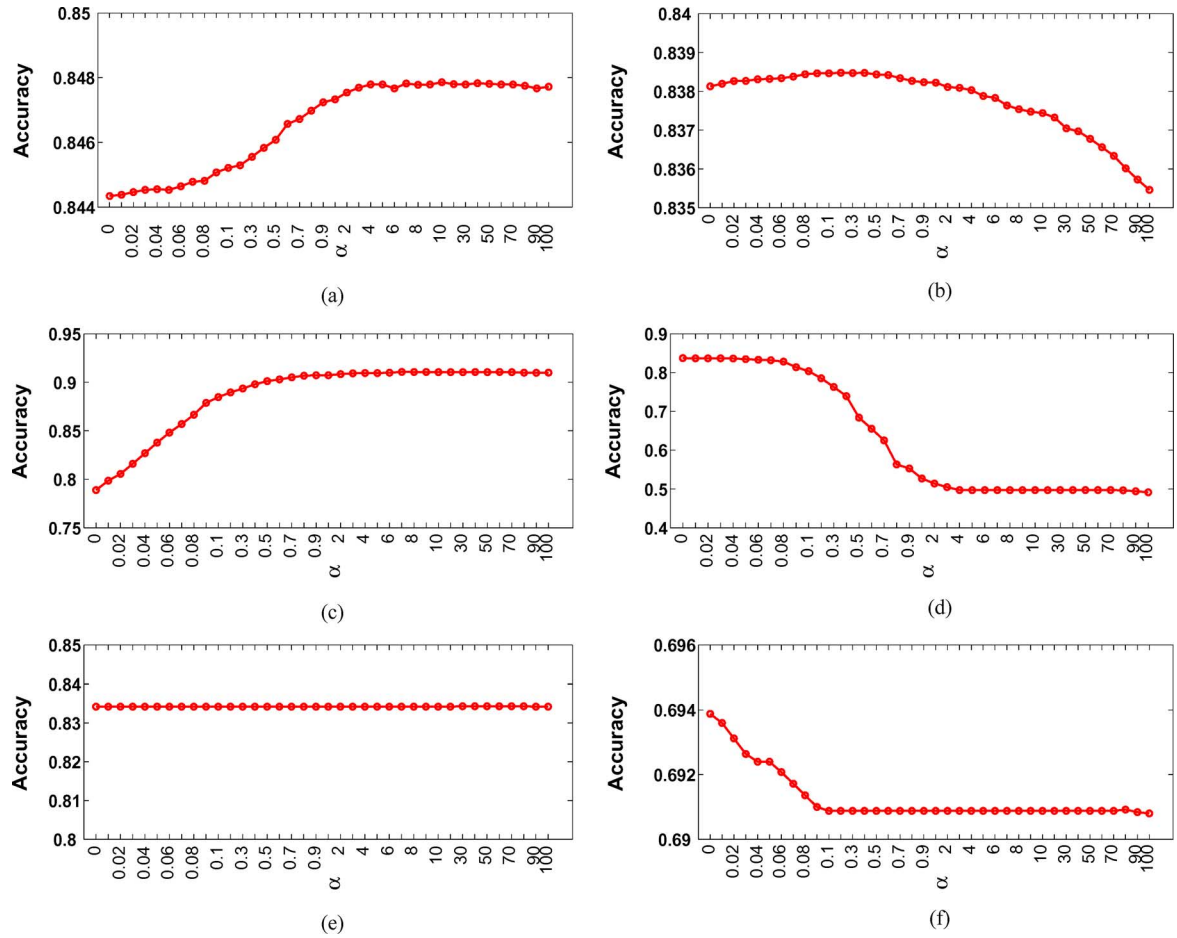


Fig. 6. Results of simulation runs for different α values. (a) Results of scenario 1, (b) Results of scenario 2, (c) Results of scenario 3, (d) Results of scenario 4, (e) Results of scenario 5, (f) Results of scenario 6.

training based super kernel fusion and logistic regression fusion method.

— For the information sources that are of different performance (scenario 1–2), the fusion results of portfolio fusion is comparable with that of other methods, and all methods outperform the logistic regression fusion method. It may be because the logistic regression fusion method is sensitive to the majority noisy information sources.

- For highly correlated information sources (scenario 4–6),
  — For the information sources that are of the similar performance (scenario 6), the portfolio fusion method generally outperforms all the other methods except the logistic regression fusion method.
  — For the information sources that are of different performance (scenario 4–5), the portfolio fusion method is comparable with other methods, and all methods outperform the logistic regression fusion method.

C. Risk Tolerance Variation

We choose different values for risk tolerance and measure the performance in the 6 scenarios. The simulation results with different α values (from very small value, i.e., 0, to very large value, i.e., 100) are shown in Fig. 6.

It can be observed that for independent information sources scenarios, moderate values can achieve good performance. For correlated information sources scenarios (more emphasis on risk), small values can achieve better performance. In general, moderate values can always achieve rather good performance. Thus, we assume a moderate appetite for risk and fix $\alpha = 2$.

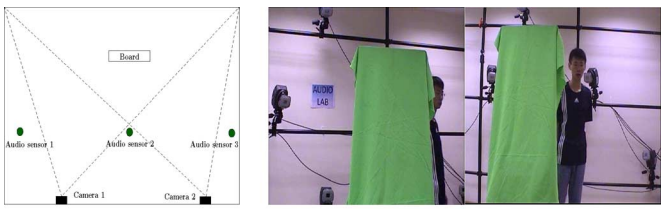
VI. Concept Detection Using Portfolio Fusion

To test the proposed portfolio fusion method for real applications, we evaluated it for concept detection on MSRA-MM dataset [25]. There are 10 000 images labeled with respect to each concept. The dataset is equally divided into development and test sets: 5000 images are selected for development of concept detection, and the other 5000 images are for testing of concept detection. In the dataset, there are 50 concepts labeled non-exclusively for the images, such as mountain, ocean, indoor, building, and cartoon.

In this experiment, four types of features from each image are exploited, including: (1) 64D HSV color histogram; (2) 256D RGB color histogram; (3) 75D edge distribution histogram; (4) 128D wavelet texture; (5) 225D block-wise color moment; (6) 144D color correlogram; (7) 7D face features [25]. The classification models are trained using the data from each information source with LIBSVM [22]. The attributes are scaled before applying SVM. When training SVM models, it is important to

TABLE II
M_A_P BY DIFFERENT FUSION METHODS

Methods	LGP	CA	LR	SKF	PTF
M_A_P	0.111	0.111	0.084	0.112	0.118



(a) (b)

Fig. 7. Experiment setup. (a) Sensor layout, (b) Example camera views.

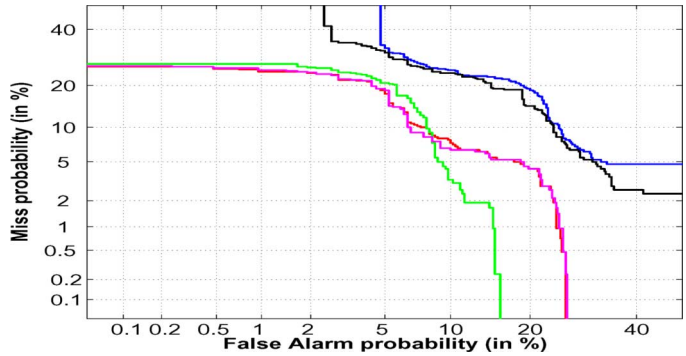


Fig. 8. DET curves of different fusion methods for human detection. According to the endpoints of the curves, from left to right, green represents LGP, red represents PTF, magenta represents CA, black represents SKF, blue represents LR.

TABLE III
DETECTION ACCURACY BY DIFFERENT FUSION METHODS

Methods	LGP	CA	LR	SKF	PTF
Accuracy	88.57%	89.29%	88.33%	87.62%	89.52%

maintain balance between the number of positive and negative samples provided [26]. In general, the concepts in the dataset are highly skewed towards negative samples (on average 6.5% positive samples). In our implementation, we utilized all available positive samples and randomly selected negative examples. In our experiments, the objective is to evaluate the relative fusion performance rather than the absolute performance. Thus, we used the RBF kernel which in general a reasonable first choice, and set the cost parameter to be 10. The other parameters were kept at default values [22].

After learning separate models for each feature, the outputs of each model are combined to obtain the fusion results. Here, we empirically set the "risk tolerance" factor $\alpha = 2$ (moderate risk) and compare our method with LGP, CA, LR, and SKF. The evaluation criteria for concept detection is the mean average precision (M_A_P), which is the mean of average precision for each concept. The average precision for each concept is calculated over the retrieved relevant images $K = 5000$. Here, only the concepts with precision larger than 0.01 in any fusion method are considered (very small average precision for all methods means the features are not good enough).

The M_A_P results are shown in Table II. It can be observed that the M_A_P of portfolio fusion method outperforms M_A_P of other fusion methods by about 5–29% (relative). It may be because our fusion method made better use of correlation and uncertainty of the decisions from different information sources.

VII. HUMAN DETECTION USING PORTFOLIO FUSION

The portfolio fusion method is also evaluated for human detection. The dataset is recorded using multiple sensors. There are three single microphone and two cameras. The sensor layout schema in Fig. 7(a) shows the relative camera and audio sensor position and overlap. The example camera views are shown in Fig. 7(b).

The task is to detect whether there is human in the region. The data is first segmented into frames and corresponding audio samples as the examples. There are 840 examples, each with one frame and corresponding audio samples for 1 second. 420 examples are selected as training set, and the remaining examples are treated as testing set. The features using for human detection are as follows:

- Audio: the audio energy. For each time interval, the audio energy can be easily calculated as the sum of squared audio samples.
- Visual: the frame different with the background. The background frame of the scene B is chosen. Then, the gray scale frame difference between any frame F and background B is calculated as $fd = F - B$.

The model for each information source is trained using LIBSVM with default parameters. The "risk tolerance" factor for portfolio fusion is empirically set to be $\alpha = 2$ here.

Table III illustrates the fusion results with different fusion methods. The audio information sources are highly correlated, and degraded the fusion performance in other fusion method. The portfolio fusion method makes use of the correlation and uncertainty. It outperforms the other methods by about 0.2–2.1% (relative). The Detection Error Tradeoff (DET) curves for different fusion methods is shown in Fig. 8. It can be observed that PTF method performs similarly to CA method, and obtains slightly better miss probability than LGP method for low false alarm. LR and SKF methods generally perform worst. It is worth mentioning that the DET curve measurement is not consistent with our return definition here. The accuracy is used as performance measure and the predicted label influences the performance instead of the actual scores. The return definition gives $+1$ return no matter what the decision score is as long as it is larger than 0.5. Thus, the actual scores do not matter as long as they return correct prediction labels. But for DET curve, the actual scores matters for the miss probability and false alarm rate. Thus, DET curve is not an appropriate measure.

VIII. CONCLUSION

In this paper, a novel multimedia fusion method using portfolio theory is proposed. The proposed method can be applied to either probabilistic output or decision output. Moreover, it

is easily scalable. Our proposed fusion method does not require additional learning for weights after models for each information source are trained. When a new information source is introduced, only the correlations will be computed instead of training the fusion model again. With well defined returns and risk, portfolio fusion method tries to maximize the return while minimizing the risk. Using appropriate definition of returns and risk, the method can also be adapted to different application scenarios. It is shown to achieve good performance in both simulation and actual experiments. The proposed fusion method can be tuned for different risk appetite of applications by using proper risk tolerance values. More study will be done on exploiting recent advances in modern portfolio theory, such as dynamic correlations adaptation, for improving the performance.

REFERENCES

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

[2] B. V. Dasarathy, *Decision Fusion*. Washington, DC: Computer Society Press, 1994.

[3] M. Stone, "The opinion pool," *Ann. Math. Statist.*, vol. 32, no. 4, pp. 1339–1342, 1961.

[4] O. Punska, "Bayesian approaches to multi-sensor data fusion," Master's thesis, Univ. Cambridge, Cambridge, U.K., 1999.

[5] J. Manyika and H. Durrant-Whyte, *Data Fusion and Sensor Management: A Decentralized Information-Theoretic Approach*. Englewood Cliffs, NJ: Prentice Hall, 1994.

[6] D. M. Tax, M. V. Breukelen, R. P. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying?," *Pattern Recognit.*, vol. 33, pp. 1475–1485, 2000.

[7] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proc. ACM Int. Conf. Multimedia*, 2004, pp. 572–579.

[8] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 226–239, 1998.

[9] T. Heskes, "Selecting weighting factors in logarithmic opinion pools," *Adv. Neural Inf. Process. Syst.*, pp. 266–272, 1998.

[10] J. Mauclair and J. Pinquier, "Fusion of descriptors for speech/music classification," in *Proc. Eur. Signal Processing Conf.*, 2004, pp. 1285–1288.

[11] R. Benmokhtar and B. Huet, "Perplexity-based evidential neural network classifier fusion using mpeg-7 low-level visual features," in *Proc. ACM Int. Conf. Multimedia Information Retrieval*, 2008, pp. 336–341.

[12] C.-W. Ngo, Y.-G. Jiang, X.-Y. Wei, F. Wang, W. Zhao, H.-K. Tan, and X. Wu, "Experimenting VIREO-374: Bag-of-visual-words and visual-based ontology for semantic video indexing and search," in *Proc. NIST TRECVID Workshop*, 2007.

[13] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, pp. 23–37, 1997.

[14] M. Li, Y. Zheng, S. Lin, Y.-D. Zhang, and T.-S. Chua, "Multimedia evidence fusion for video concept detection via OWA operator," in *Proc. Int. Conf. MultiMedia Modeling*, 2009, vol. 5371, pp. 208–216.

[15] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, "Aggregation for Gaussian regression," *Ann. Statist.*, vol. 35, no. 4, pp. 1674–1697, 2007.

[16] A. Juditsky and A. Nemirovski, "Functional aggregation for nonparametric regression," *Ann. Statist.*, vol. 28, no. 3, pp. 681–712, 2000.

[17] X. Wang and M. S. Kankanhalli, "Portfolio theory of multimedia fusion," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 723–726.

[18] H. Markowitz, "Portfolio selection," *J. Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[19] A. Černý, *Mathematical Techniques in Finance: Tools for Incomplete Markets*. Princeton, NJ: Princeton Univ. Press, 2003.

[20] L. Breiman, Bias, Variance, and Arcing Classifiers, University of California Berkeley, TechReport 460, 1996.

[21] N. Poh and S. Bengio, "How do correlation and variance of base-experts affect fusion in biometric authentication tasks?," *IEEE Trans. Signal Process.*, vol. 53, no. 11, pp. 4384–4396, Nov. 2005.

[22] C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[23] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 1999, pp. 61–74.

[24] R. Aly and D. Hiemstra, "Concept detectors: How good is good enough?," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 233–242.

[25] M. Wang, L. Yang, and X.-S. Hua, Msra-mm: Bridging Research and Industrial Societies for Multimedia Information Retrieval, Microsoft Research (MSR), TechReport MSR-TR-2009-30, 2009.

[26] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu, Columbia University's Baseline Detectors for 374 Lscom Semantic Visual Concepts Columbia University, TechReport 222-2006-8, 2007.

Xiangyu Wang received the B.Sc. degree in computer science and technology from Fudan University. He is pursuing the Ph.D. degree at the School of Computing, National University of Singapore. His current research interests are in multimedia data analysis, retrieval, and modeling. He has authored/co-authored several research articles in reputed ACM, IEEE and Springer journals and conferences. Mr. Wang has been a Reviewer for various journals and conferences.

Mohan Kankanhalli is a Professor at the Department of Computer Science of the National University of Singapore. He is also the Associate Provost for Graduate Education at NUS. He obtained his B.Tech. (electrical engineering) from the Indian Institute of Technology, Kharagpur, in 1986 and his M.S. and Ph.D. (computer and systems engineering) from the Rensselaer Polytechnic Institute in 1998 and 1990, respectively. He is actively involved in the organization of many major conferences in the area of Multimedia. He is on the editorial boards of several journals including the ACM Transactions on Multimedia Computing, Communications, and Applications, Springer Multimedia Systems Journal, Pattern Recognition Journal and Multimedia Tools & Applications. His current research interests are in Multimedia Systems, Digital Video Processing and Multimedia Security (surveillance, digital rights management and privacy).