Multimedia Content Processing through Cross-Modal Association

Dongge Li[1]
[1]Motorola Labs
1301 East Algonquin Road
Schaumburg, Illinois 60196
dongge.li@motorola.com

Nevenka Dimitrova[2]
[2]Philips Research
345 Scarborough Rd.
Briarcliff Manor, NY 10510
nevenka.dimitrova@philips.com

Mingkun Li[3], Ishwar K. Sethi[3]
[3]Intelligent Info. Eng. Lab
Oakland University
Rochester, MI 48309
{li,isethi}@oakland.edu

ABSTRACT
Multimodal information processing has received considerable attention in recent years. The focus of existing research in this area has been predominantly on the use of fusion technology. In this paper, we suggest that *cross-modal association* can provide a new set of powerful solutions in this area. We investigate different cross-modal association methods using the linear correlation model. We also introduce a novel method for cross-modal association called Cross-modal Factor Analysis (CFA). Our earlier work on Latent Semantic Indexing (LSI) is extended for applications that use off-line supervised training. As a promising research direction and practical application of cross-modal association, cross-modal information retrieval where queries from one modality are used to search for content in another modality using low-level features is then discussed in detail. Different association methods are tested and compared using the proposed cross-modal retrieval system. All these methods achieve significant dimensionality reduction. Among them CFA gives the best retrieval performance. Finally, this paper addresses the use of cross-modal association to detect talking heads. The CFA method achieves 91.1% detection accuracy, while LSI and Canonical Correlation Analysis (CCA) achieve 66.1% and 73.9% accuracy, respectively. As shown by experiments, cross-modal association provides many useful benefits, such as robust noise resistance and effective feature selection. Compared to CCA and LSI, the proposed CFA shows several advantages in analysis performance and feature usage. Its capability in feature selection and noise resistance also makes CFA a promising tool for many multimedia analysis applications.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Algorithms, Indexing methods, Video analysis*

General Terms
Algorithms, Measurement, Theory.

MM'03, November 2-8, 2003, Berkeley, California, USA.

Keywords
Cross-modal association, cross-modal information retrieval, talking head analysis, cross-modal factor analysis (CFA)

1. INTRODUCTION
Video content usually contains events with synchronized audio and visual elements. Both aspects carry their contribution to the high level semantics, and the presence of one has usually a "priming" effect on the other: when hearing a dog barking we expect the image of a dog, seeing a talking face we expect the presence of her voice, image of a waterfall usually bring the sound of running water etc. A series of psychological experiments on cross-modality influence [1] have proved the importance of synergistic integration of multiple modalities in the human perception system. A typical example of this kind is the well-known McGurk effect [2]. Yet, in video content analysis and retrieval, video compression, visual speech recognition, even talking head animation, the analysis is usually performed separately on the different modalities: first on the visual signal, then the audio signal and the results are brought together using fusion methods. However, in this process of separation of modalities, we lose valuable information about the whole event and/or object we are trying to analyze and detect. There is an inherent association between the two modalities and the analysis can take advantage of the synchronized appearance of the relationship between audio and visual signal.

This paper will show the potential of cross-modality information analysis methods, which extract multimedia content by identifying and measuring intrinsic associations between different modalities. We refer to such approaches of extracting multimedia content as 'analysis by semantic association' or 'cross-modal association'. This paper will present several cross-modal association approaches under the linear correlation model: the latent semantic indexing (LSI), canonical correlation analysis (CCA) and Cross-modal Factor Analysis (CFA). LSI and CCA have been explored on the topic of talking head detection. In this paper, we will generalize the applications of these methods to capture associations across modalities in synchronized audiovisual signal. We will propose the Cross-modal Factor Analysis as a solution. These proposed methods are compared for cross-modal information retrieval and talking head analysis.

Depending on the nature of applications, different kinds of methods could be used to identify and measure semantic associations between different modalities. In general, two types of approaches are expected: (i) model-based approaches, which employ certain

association models like Gaussian distribution or linear correlation models, and (ii) model-free approaches like neural networks based approaches. Model-free approaches require little prior knowledge and are pertinent to more general applications. However, if appropriate models are applied, model-based approaches require less training examples and can generate better results. For many audiovisual data analysis applications, if the analysis time window can be relatively short, linear relationship could be an appropriate model between synchronized audio and visual features [3, 4, 5]. This paper will mainly focus on cross-modal association approaches under the linear correlation model.

In [6] Hershey and Movellan proposed the use of mutual information evaluation to measure the synchrony between audio and visual signal. Their mutual information estimation is calculated under the assumption of Gaussian distribution of audio and visual features. Slaney and Covell [5] presented FaceSync as an optimal linear detector, which combines the information from all the pixels to measure audio-visual synchronization. Their approach is based on two surprisingly simple algorithms in computer-vision and audio-visual speech synthesis: EigenPoints [3] and multilinear facial synthesizer [7]. EigenPoints [3] is an algorithm that finds a linear mapping between the brightness of a video signal and the location of fiduciary points on the face. The EigenPoints can find a linear approximation that describes the brightness–fiduciary space, and this linear approximation is valid over a useful range of brightness and control-point changes. Yehia et al. [7] have presented a method to connect a specific model of speech, the line-spectral pairs or LSP, with the position of fiduciary points on the face. Their multilinear approximation yielded an average correlation of 0.91 between the true facial locations and those estimated from the audio data. In [9], Fisher et al. presented a non-parametric approach to learn the joint distribution of audio and visual features. They first project the data into a maximally informative, low-dimensional subspace, and then model the stochastic relationships using a nonparametric density estimator. Iyenger et al. investigated monologue detection using mutual information model [10]. They introduced two techniques as VQ-based MI and Gaussian-based MI respectively. With either scheme, the face amongst a set of possibilities that is deemed to have produced a given audio sequence provides the highest mutual information score. We introduced LSI method for face-speech matching [4] and compared it to the correlation method. We discovered that when the two methods are contrasted, the LSI method has overall better performance, as well as graceful degradation in presence of noise.

Information retrieval via cross-modal querying has many applications. Popular web search engines including Google and Yahoo have already started to provide cross-modal retrieval functionality by retrieving images based on textual information in image titles or their associated documents. Several research groups have also addressed the use of language-related information (such as keywords, syllables, etc.) to link different modalities [12, 13]. While simple and straightforward, these initial efforts have some major limitations: (a) They depend essentially on the robust generation of high-level descriptions from multimedia data. However, bts of these tasks, such as speech-to-text recognition, object recognition, and video summary, are generally recognized as very challenging research topics and have so far achieved only limited success in controlled conditions. Even with good algorithms, it may still be very difficult to obtain robust results in adverse environments for many sophisticated recognition tasks. (b) There are plenty of cross-modal retrieval applications involving multimedia content/features that cannot be represented by high-level descriptions. Like traditional content-based retrieval problems, low-level features in many applications can be very important. However, existing systems based on high-level descriptions cannot deal with most of such features. A complete cross-modal retrieval system should be able to handle features at different levels. The retrieval procedure could involve various features in different modalities. For example, a video clip may match with certain background music due to both high-level content such as events and low-level features such as tempo and dominant color. In this paper, we will discuss the use of cross-modal association for information retrieval, which provides new solutions to address the above essential issues. Cross-modal association offers effective ways to associate features at various levels across modalities. We hope that efforts in this area will lead to the development of better and more comprehensive multimodal information retrieval systems.

The organization of this paper is as follows: In Section 2 we present latent semantic indexing approach, in Section 3 we describe cross-modal factor analysis, and in Section 4 we present canonical correlation analysis. In Section 5 we discuss two applications: cross-model information retrieval and talking head detection. In Section 6, we compare the propose methods based on some experimental results. Conclusions and further work suggestions are given in Section 7.

2. LATENT SEMANTIC INDEXING

LSI is used as a powerful tool in text information retrieval to discover underlying semantic relationship between different textual units (e.g. keywords and paragraphs). In [4] we propose a method to detect the semantic correlation between visual faces and its associated speech based on LSI. This method consists of four steps: the construction of a joint multimodal feature space, normalization, singular value decomposition (SVD), and semantic association measurement.

Given n visual features and m audio features at each of the t video frames, the joint feature space can be expressed as:

$$X = [V_1, V_2, ..., V_n, A_1, A_2, ..., A_m] \tag{1}$$

where

$$V_i = (v_i(1), v_i(2), ..., v_i(t))^T \tag{2}$$

and

$$A_i = (a_i(1), a_i(2), ..., a_i(t))^T \tag{3}$$

Various visual and audio features can have quite different variations. Normalization of each feature in the joint space according to its maximum elements (or certain other statistical measurements) is thus needed and can be expressed as:

$$\hat{X}_i(:) = \frac{X_i(:)}{\max(abs(X_i(:)))} \tag{4}$$

After normalization all elements in normalized matrix $\hat{X}$ have values between –1 and 1. SVD can then be performed as follows:

$$\hat{X} = S \cdot V \cdot D^T \tag{5}$$

where S and D are matrices composing of left and right singular vectors and V is the diagonal matrix of singular values in descending order.

Keeping only the first and most important k singular vectors in S and D, we can derive an optimal approximation of $\hat{X}$ with reduced feature dimensions, where semantic (correlation) information between visual and audio features is mostly preserved and irrelevant noise is greatly reduced. Traditional Pearson correlation or mutual information calculation [4, 6, 9] can then be used to effectively identify and measure semantic associations between different modalities. Experiments in [4] have shown the effectiveness of LSI and its advantages over the direct use of traditional correlation calculation.

The above optimization of $\hat{X}$ in the least square sense can be expressed as:

$$\hat{X} \cong \tilde{X} = \tilde{S} \cdot \tilde{V} \cdot \tilde{D}^T \tag{6}$$

where $\tilde{S}$, $\tilde{V}$, and $\tilde{D}$ consist of the first k vectors in S, V, and D, respectively. The selection of an appropriate value for k is still an open issue in the literature. In general, k has to be large enough to keep most of the semantic structures and small enough to remove some irrelevant noise.

Equation (6) is not applicable for applications using off-line training since the optimization has to be performed on the fly directly based on the input data. However, due to the orthogonal property of singular vectors, we can rewrite (6) in a new form as follows:

$$\hat{X} \cong \tilde{X} = X \cdot \tilde{D} \cdot \tilde{D}^T \tag{7}$$

Now we only need the $\tilde{D}$ matrix in the calculation, which can be trained in advance using groundtruth data. This derived new form is important for those applications that need off-line trained SVD results. One of such examples is cross-modal information retrieval presented in Section 5.1. We will later use and test this new form in Section 6 of this paper.

3. CROSS-MODAL FACTOR ANALYSIS (CFA)

LSI does not distinguish features from different modalities in the joint space. However the optimal solution based on overall distribution may not best represent semantic relationships between features of different modalities, since distribution patterns among features from the same modality will also greatly impact LSI's results. This can be shown in Figure 1, where there is a big difference between the principal distribution direction indicated by the first vector of D matrix in LSI and the actual correlated direction between visual features $\{v_1, v_2\}$ and audio feature a_1.

A solution to the above problem is to treat features from different modalities as two subsets and focus only on semantic patterns between these two subsets. Distribution patterns and noise within each subset should not be a distraction factor (as did for LSI results shown in Figure 1). Under the linear correlation model, the problem now is to find the optimal transformations that can best represent (or identify) the coupled patterns between features of two different subsets. We adopt the following optimization criterion in the rest of this section:

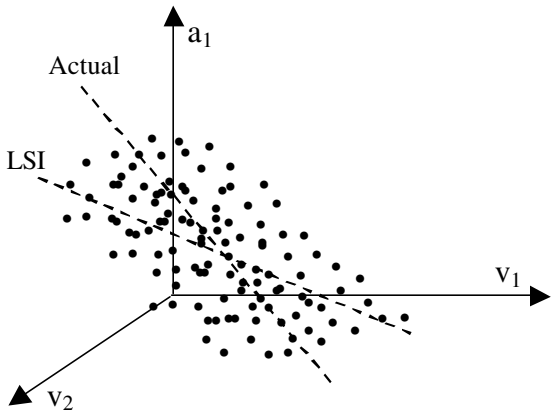

Figure 1. Illustration of principal coupled direction between visual features $\{v_1, v_2\}$ and audio feature a_1

Given two mean centered matrices X and Y, which compose of row-by-row coupled samples from two subsets of features, we want orthogonal transformation matrices A and B that can minimize the expression:

$$\left\| XA - YB \right\|_F^2 \tag{8}$$

where $A^T A = I$ and $B^T B = I$. $\left\| M \right\|_F$ denotes the Frobenius norm of the matrix M and can be expressed as:

$$\left\| M \right\|_F = \left(\sum_i \sum_j \left| m_{ij} \right|^2 \right)^{1/2} \tag{9}$$

In other words, A and B define two orthogonal transformation spaces where coupled data in X and Y can be projected as close to each other as possible.

Since we have:

$$\begin{aligned} \left\| XA - YB \right\|_F^2 &= trace\left(\left(XA - YB \right) \cdot \left(XA - YB \right)^T \right) \\ &= trace(XAA^T X^T + YBB^T Y^T - XAB^T Y^T - YBA^T X^T) \\ &= trace(XX^T) + trace(YY^T) - 2 \cdot trace(XAB^T Y^T) \end{aligned} \tag{10}$$

where trace of a matrix is defined to be the sum of the diagonal elements. We can easily see from above that matrices A and B which maximize $trace(XAB^T Y^T)$ will minimize (8). It can be shown [15] that such matrices are given by:

$$\begin{cases} A = S_{xy} \\ B = D_{xy} \end{cases} \quad \text{where } X^T Y = S_{xy} \cdot V_{xy} \cdot D_{xy} \tag{11}$$

With the optimal transformation matrices A and B, we can calculate the transformed version of X and Y as follows:

$$\begin{cases} \tilde{X} = X \cdot A \\ \tilde{Y} = Y \cdot B \end{cases} \tag{12}$$

Corresponding vectors in $\tilde{X}$ and $\tilde{Y}$ are thus optimized to represent the coupled relationships between the two feature subsets without being affected by distribution patterns within each subset. Traditional Pearson correlation or mutual information calculation [4,

6, 9] can then be performed on the first and most important k corresponding vectors in $\tilde{X}$ and $\tilde{Y}$, which similarly to those in LSI preserve the principal coupled patterns in much lower dimensions and at the same time remove irrelevant noise.

In addition to feature dimension reduction and noise removal, feature selection capability is another feature of CFA. The weights (or loadings) in A and B automatically reflect the significance of individual features. We show in Figure 2 the absolute values of the first seven vectors of matrix A obtained from the training of 300 frames of visual faces and their associated speech features. For better visualization each vector has been reshaped according to the corresponding visual location. It is obvious that matrix A is able to highlight those facial areas corresponding most to the speech. This clearly demonstrates the great feature selection capability of CFA, which makes it a promising tool for many multimedia analysis applications, including multimodal face localization, audiovisual speech recognition, multimodal noise cancellation, etc.

Figure 2. Absolute values of the first seven vectors of matrix A reshaped according to the corresponding visual location.

4. CANONICAL CORRELATION ANALYSIS (CCA)

Following the development of the previous section, we can adopt a different optimization criterion: Instead of minimizing the projected distance, we attempt to find transformation matrices A and B that maximize the correlation between XA and YB. This can be described more specifically using the following mathematical formulations:

Given two mean centered matrices X and Y as defined in the previous section, we seek matrices A and B such that

$$correlation(XA, YB) = correlation(\tilde{X}, \tilde{Y}) = diag\{\sigma_1, \sigma_2, \cdots, \sigma_l\} \tag{13}$$

where $\tilde{X} = X \cdot A$, $\tilde{Y} = Y \cdot B$, and $1 \geq \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_l \geq 0$. σ_i represents the largest possible correlation between the ith translated features in $\tilde{X}$ and $\tilde{Y}$. Note that A and B are only determined up to a constant factor by (13). A statistical method called canonical correlation analysis [16, 17] can solve the above problem with additional norm and orthogonal constraints on translated features:

$$E\{\tilde{X}^T \cdot \tilde{X}\} = I \text{ and } E\{\tilde{Y}^T \cdot \tilde{Y}\} = I \tag{14}$$

In CCA, A and B are calculated as follows:

$$A = \Sigma_{xx}^{-1/2} \cdot S_K \text{ and } B = \Sigma_{yy}^{-1/2} \cdot D_K \tag{15}$$

where

$$\Sigma_{xx} = E\{X^T \cdot X\}$$

$$\Sigma_{yy} = E\{Y^T \cdot Y\}$$

$$\Sigma_{xy} = E\{X^T \cdot Y\}$$

and

$$K = \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} = S_K \cdot V_K \cdot {D_K}^T \tag{16}$$

The calculation of inverse matrix requires that no linear correlation exists between any two vectors within X or Y. Large calculation errors could result even when two vectors are just close to linear. This imposes some restrictions on the set of features that can be processed by CCA, especially when the analysis window has to be relatively short to fit the linear model across modalities. As we will see later in our experiments, such restriction of CCA may sometimes affect its performance and limit its applications. CFA proposed in the last section, however, does not have such restrictions. It can directly process whatever best feature sets available.

Other major differences between CCA and CFA include:

(a) The transformations provided by CFA are orthogonal, while this is not necessary true for CCA. It can be seen from Equation (11) that A and B given by CFA satisfy $A^TA=I$ and $B^TB=I$, where I is the identity matrix. CCA, however, does not provide such orthogonal transformations in most cases.

(b) CFA is in favor of coupled patterns with high variations (i.e. large amplitude changes), while CCA is more sensitive to highly coupled, but low variation patterns. This is mainly due to the whitening of X and Y in CCA by calculating $\Sigma_{xx}^{-1/2}$ and $\Sigma_{yy}^{-1/2}$.

5. APPLICATIONS

Discovering the inherent associations between audio and visual aspects of the video signal has many potential applications. These applications include content-based retrieval of multimedia data, audiovisual content analysis, audiovisual compression, facial animation, audiovisual speech recognition, videoconferencing, video editing (e.g. synchrony of audio and video content), etc. In this paper we will focus our discussion on the content-based retrieval and audiovisual content analysis applications.

5.1 Cross-Modal Information Retrieval

With cross-modal association, heterogeneous features extracted from different media sources (e.g. audio and images) can be matched against each other based on the synchronized correlation patterns. This enables a new practical application – cross-modal retrieval, where query from one type of media sources (e.g. audio) can be used to search for content on a different type of media sources (e.g. image sequences). The search can be based directly on *low-level features*, which is similar to traditional content-based multimedia retrieval.

One application for cross-modal retrieval is to compensate corrupted (or absent) media sources. For example, if a sound is corrupted by background noise (or absent), an associated visual feature can be used, instead, as the basis for search. It is possible to further integrate cross-modal methods with existing single-modal retrieval methods to provide robustness and capabilities not possessed when using either of the media types alone.

The methods also offer the user greater choice in browsing a multimedia database because the user can select the modality she or he prefers and with which the user is most familiar. This approach also has the advantage of enabling the user to browse and search multimedia content of different modalities in a manner that

minimizes bandwidth. For example, instead of passing a query in the form of an image over a network, for example the Internet, only a voice query needs to be transmitted to retrieve an image. This approach offers multiple choices for input modalities: we can choose a microphone as an input device instead of a graphic input device.

Figure 3 shows the modular structure of an audio-visual retrieval system, where queries of speech are used to search for image sequences with similar motions for the speech. Note that the search is performed without the presence of sound tracks associated with image sequences.

In Figure 3, the visual features used for cross-modal retrieval are simply aligned pixel intensities or eigenface values from candidate face areas. An omni-face detection system capable of locating frontal- and side-view faces is used to generate candidate face areas [14]. All the candidate face areas are then scaled to areas of 40x32 pixels, based on which low-level visual features are generated. The audio features are 12 Mel-Frequency Cepstral Coefficients (MFCCs), which have been widely used in speech recognition and speaker identification applications. In our implementation the MFCCs are extracted using the DCT of filter-banked FFT spectra.

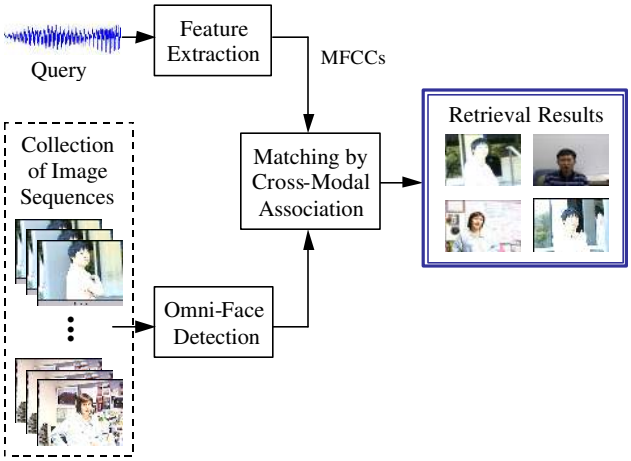

Figure 3. Cross-Modal Retrieval System Modules.

Depending on the association approach chosen, an off-line supervised training process can be performed based on Equation (5), (11), or (15) to calculate optimized transformation matrices for the low-level audio and visual features. Video clips with talking faces and a synchronized speech are used as groundtruth data for the training. Once the process completed, the training results can then be used for the retrieval process.

According to cross-modal association methods discussed in previous sections, the transformation matrices are used to transform the low-level features into a reduced feature space, where the matching between query and search candidates on different types of media sources can be evaluated. For CFA and CCA the transformation is based on Equation (12). For LSI the derived new transformation form Equation (7) is needed. The evaluation of matching can then be performed based on Person correlation or mutual information in the transformed space. In our implementation Pearson correlation is chosen for its simplicity. Candidates with highest Pearson correlations are considered as the best matches.

Figure 4 shows an example of results generated by our cross-modal retrieval system. The query is a 0.3 second audio clip containing the speech syllable /ke/. The search is performed on a total of over 300 candidate short sequences corresponding to different speech syllables. We show in Figure 4 the top six retrieval results, according to the descending order of CFA correlation values. To better compare the retrieved face statuses, the results contain only the face areas in the images with their corresponding syllables. The actual faces are much smaller in the original image sequences as shown in the lower-left corner of the figure. While many different face statuses are included in the test, as can be seen from the retrieval results all the faces retrieved are in very similar status. In particular, the second and sixth results correspond to exact the same syllable as the audio query.

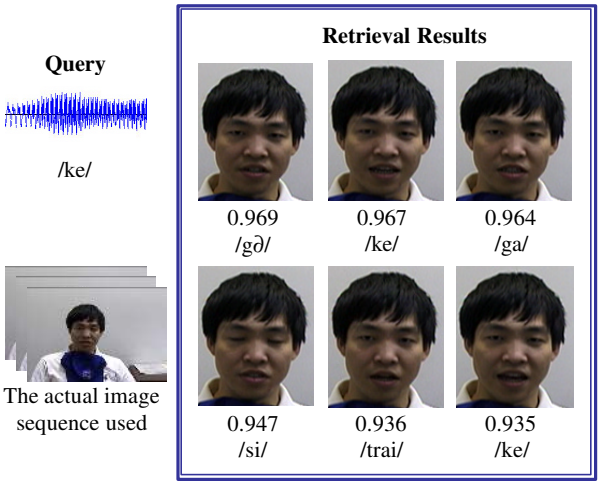

Figure 4. An example of cross-modal retrieval results.

All three cross-modal association methods proposed earlier have been tested and compared for cross-modal retrieval. According to our experimental results, fewer than ten most important vectors within the transformation matrices are normally needed to achieve the good performance, which is a significant reduction of feature dimension after the optimized transformation. Before the transformation we have 12 audio features and 1280 visual features when pixel intensities are used. A good feature dimension reduction will greatly save storage and improve searching speed. We will further discuss our experimental results in Section 6.

5.2 Talking Head Analysis

Audio-visual content analysis is another important application area for cross-modal association methods. This section will in particular focus on talking head analysis, an emerging research topic in audio-visual analysis.

Talking head analysis is the process to automatically detect the face(s) on the screen, if any, that has corresponding speech in the synchronized soundtrack. Talking head analysis is important for video indexing, videoconferencing, and person identification. Using talking head analysis we can classify shots into different categories according to cast presence and roles in narrative content. Existing audio-visual person identification systems [8] require that the

speaking face is the only face appears on the screen. Such a requirement, which greatly limits their usage in general applications, can be removed with the talking head analysis capability.

Figure 5 shows the block diagram of our talking head analysis system. The core part of this system is the audiovisual cross-modal association module, where coupled patterns between low-level visual features and audio features are trained and evaluated using methods discussed earlier. The visual features and audio features are the same as those used in cross-modal retrieval. Again the omni-face detection system is used to locate frontal- and side-view face areas [14].

A notable feature of our system is the use of audio classification to trigger cross-modal association. Since faces can be in any status when not speaking or during the pause of a speech, it will be better to exclude audio and visual features within those periods from training or evaluation. We developed an audio classification system that can robustly segment and classify general audio data into seven categories including music, noise, silence, speech, speech+music, speech+noise and speech+music. Cross-modal assoication will only be perfromed during audio segments of the latter four categories, which have speech components. A detailed discussion of our audio classification algorithm can be found in [18].

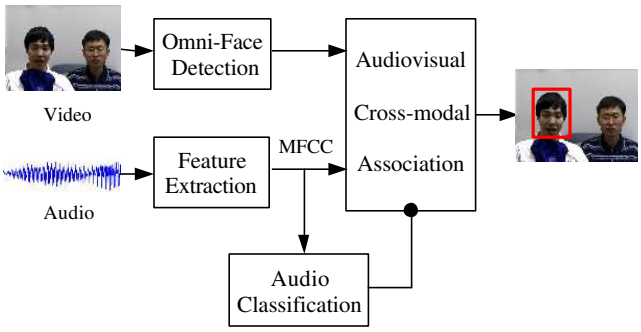

Figure 5. Talking head analysis system structure.

Based on these low-level audio and visual features, we can calculate optimized transformation matrices using cross-modal association methods discussed in previous sections. The calculation can be performed either offline using some 'groundtruth' training data or on-the-fly using the input data directly. We refer to the first case as off-line supervised training and the second case as on-line dynamic calculation of transformation matrices. In the latter case, the features from candidates are not necessary to be the right corresponding ones. Only the correct corresponding features with certain correlation patterns will provide good match in the later evaluation step.

With the matrices generated, the audio and visual features can be transformed into the new feature space. Pearson correlation or mutual information is then calculated in the transformed feature space for each face areas. Whenever multiple candidates exist, the real talking head is detected as the one with highest Pearson correlation or mutual information in the transformed feature space. We will compare different cross-modal association methods for the talking head analysis system in the next section.

6. EXPERIMENTAL RESULTS

A series of experiments have been conducted to test and compare the proposed association methods for cross-modal information retrieval and talking head analysis problems. We used different types of video material, including nine home video clips, six movie clips, and three video conferencing clips, all of which are dialog clips with multiple people speaking. The total duration of video in the experimental data set is about 30 minutes. While some video clips are in good quality with close-up faces and low audio noise, others have small faces and/or high noise level. Faces in many video clips are in very natural positions and not necessary facing the camera.

In cross-modal retrieval experiments, we use four-second speech clips as queries to search for image sequences corresponding to the speech. The performance is evaluated by the rank of the correct 'hit' in the retrieval list. Table 1 gives the percentage of correct hits appears in the top n of the retrieval lists. Among the three methods, CFA and CCA are much better than LSI in all cases. This demonstrates the effectiveness of optimization provided by CFA and CCA. CFA also slightly outperforms CCA in most cases. More than half of the correct hits using CFA are within the top 10 retrieval results.

	CFA	CCA	LSI
Top 1	20%	15%	0%
Top 5	30%	25%	5%
Top 10	55%	50%	10%
Top 20	80%	80%	30%

Table 1. Comparison of different cross-modal association methods in cross-modal retrieval.

In Figure 6 we show the CFA correlation values between the audio query and different locations on an image sequence. The circled dot indicates the actual image frame location corresponding to the audio query. It is obvious that the corresponding images are correctly identified in the retrieval results with apparently the highest CFA correlation values. In addition, we can see from the graph that the next two best retrieval results have also very similar visual appearance. Those images with low CFA correlation values are in quite different status. This, in addition to the results shown in Figure 4 of Section 5.1, illustrates the effectiveness of CFA method for cross-modal retrieval.

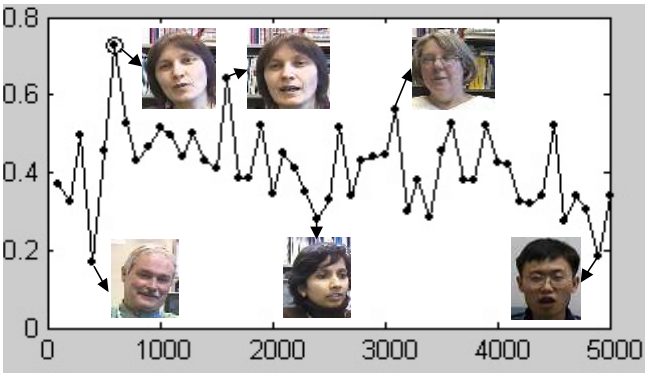

Figure 6. CFA correlation values at different locations on an image sequence.

With another set of experiments, we further tested our methods for the retrieval of explosion image sequences using audio queries. The experiments are conducted using a separate data collection, which includes 452 explosion clips and 3870 non-explosion clips collected from the movie Matrix, 9/11 live broadcasts, and the Internet. Many of the 9/11 clips are recorded in poor quality with no soundtrack. Again, 12 MFCCs are used as audio features. The visual features are obtained through three processing steps. We first partition each image into 5x10 overlapped blocks. For each image block three local histograms can then be calculated in the HSI color space. Finally, an *area-peak* value is extracted from each local histogram through an operation very similar to low-pass filtering. We thus have a total of 150 *area-peak* values representing each image. Details on the calculation of *area-peak* values can be found in [11]. Figure 7 shows an example image and its representation using 150 *area-peak* values.

Figure 7. An example image and its representation using 150 area-peak values.

Using 70 explosion video clips and their associated audio, we trained all three cross-modal association methods. The tests are then performed on the rest of data using 3-second audio queries to search for explosion image sequences, many of which have no soundtrack. The performance is evaluated according to the percentage of correct hits in the ranked retrieval list. Table 2 illustrates the performance of different association methods. The results are consistent with our previous tests. CFA and CCA again greatly outperform LSI in all cases. The difference between CFA and CCA is marginal. In Figure 8, we give a typical example of retrieval results using CFA method.

One limitation of our current system is the lack of tolerance of time variation in the matching process. A short visual event will be missed if a similar event given by the audio query is much longer. This can severely degrade the performance especially as the ranked retrieval list goes longer. However, the problem can be fixed by using dynamic matching algorithms such as dynamic time wrapping (DTW). A good matching method will thus further improve the retrieval performance.

	CFA	**CCA**	**LSI**
Top 5	62%	61%	21%
Top 10	41%	42%	21%
Top 20	37%	32%	20%

Table 2. Comparison of different cross-modal association methods for the retrieval of explosion images.

In the above experiments on cross-modal retrieval, only 8 most important feature dimensions in the transformed feature space are used. In other words, we only need 16 features per image, 8 of which are visual features and the rest are corresponding audio features. Comparing to the original data volume and feature dimension size, there is a significant save in space for the database. This further shows the effectiveness of the given methods in discovering and representing cross-modal information. In addition, the highly compact representation can greatly speed-up the search/retrieval process and facilitate many data management tasks.

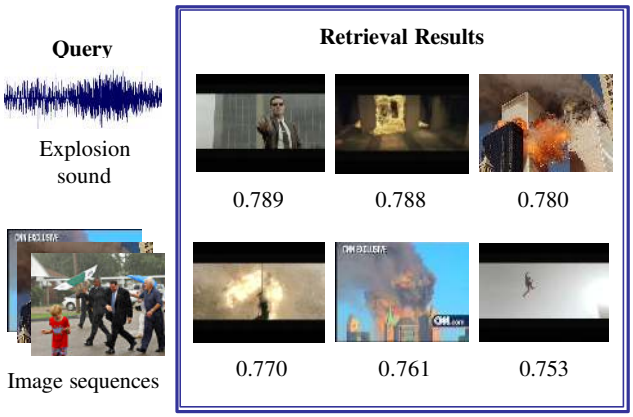

Figure 8. Cross-modal retrieval of explosion image sequences using FCA.

In our talking head analysis experiments, the real talking head competes either with other faces in the video or talking faces chosen from other image sequences in the corpus. Two different types of processing are used for each of the methods discussed earlier: off-line supervised training where the transformation matrices are generated before hand using groundtruth data, and on-line dynamic processing where the transformation matrices are generated on the fly using the input testing video directly.

Table 3 provides the performance comparison of different methods. Overall, CFA achieves 91.1% accuracy using supervised training and 80.4% accuracy using dynamic processing. CCA and LSI provide less than 75% accuracy in all tests. There is a significant improvement using CFA in both dynamic processing and supervised training cases. The success of CFA proves the effectiveness of its semantic association capability. However, CCA that follows a similar idea achieves much lower accuracy. An examination of the weights in transformation matrices (A and B) explains the difference: while CFA is able to generates meaningful patterns in A and B as shown in Figure 2, CCA provides more 'noisy' patterns in A and B that are strongly influenced by low-variation patterns. This makes CCA less tolerance to noise than CFA. As mentioned at the end of Section 4, CFA, unlike CCA, is in favor of coupled patterns with high variations. CFA thus provides better noise tolerance in talking head analysis by favoring coupled patterns with high variations.

	LSI	CFA	CCA
Dynamic processing	66.1%	80.4%	73.9%
Supervised training	53.6	91.1%	70.8%

Table 3. Accuracy of different talking head analysis methods.

For both cross-modal retrieval and talking head analysis, the feature used for CFA can be either aligned pixel intensities or eigenface values. CCA can only use eigenface values as input due to its limitation mentioned in Section 4. Our experiments on CFA method show that the use of eigenfaces slightly degrades the performance

by 3-5% in accuracy. Our guess for the reason is: the use of eigenface transformation, compared to original pixel intensities, will slightly lose some useful information needed for cross-modal association. The limitation of CCA may thus affect its performance in some applications.

7. CONCLUSIONS

Existing research in multimodal information processing has been predominantly focusing on the use of fusion technology. In this paper, however, we demonstrated that cross-modal association could also provide a set of powerful solutions in this area. New practical applications (e.g. cross-modal retrieval) and better approaches can be offered by analyzing the inherent associations between different modalities.

We systemically investigated different cross-modal association methods under the linear correlation model. Our earlier work on LSI is extended for applications that need off-line supervised training. We also proposed a novel method for cross-modal association called Cross-modal Factor Analysis. As have shown by our experiments in different applications, CFA provides a powerful tool for many potential applications of cross-modal association.

This paper also proposed cross-modal retrieval as a promising research area and practical application of cross-modal association. Cross-modal retrieval has the advantage of compensating for corrupted (or absent) media sources. It also offers the user greater choice in browsing a multimedia database. A cross-modal retrieval system is implemented in this work. We tested and compared different association methods for the retrieval of speaking faces and explosion scenes via cross-modal querying. All three methods achieved significant dimensionality reduction. Among them CFA gives the best retrieval performance. Our results demonstrated the effectiveness of cross-modal association in 'connecting' heterogeneous features across modalities. We hope that continuing efforts in this area will bring about valuable advancements in multimodal information retrieval technology.

Finally, we discussed another application for cross-modal association: talking head analysis, an emerging research topic in audio-visual analysis. We presented a talking head analysis system and compared different cross-modal association methods. CFA method achieves 91.1% detection accuracy, while LSI and Canonical Correlation Analysis (CCA) achieve 66.1% and 73.9% accuracy, respectively.

As shown by experiments in cross-modal retrieval and talking head analysis, CFA provides a powerful tool to analyze semantic associations between different modalities. Compared to CCA, CFA provides better noise tolerance capabilities and has no constraints on the features to be processed. Its capability in feature selection and noise resistance also makes CFA a promising tool for many multimedia analysis applications.

Future research in this area can be developed in many different directions. We are currently investigating the use of cross-modal association methods for several other applications. These include audiovisual speech recognition, cross-modal compression, and audio-visual animation. Taking audiovisual speech recognition as an example, the recognition performance can be improved by taking advantage of CFA's feature selection to pick frequency channels that corresponding best to speaking faces and discard noisy channels. Also, there is a lot of research work to be done in cross-modal retrieval. More sophisticated cross-modal association methods may be proposed for future systems. Existing dynamic matching algorithms such as DTW can be integrated with our methods to better handle variations in time and further improve retrieval performance.

8. ACKNOWLEDGMENTS

We would like to thank Hong-Jiang Zhang, Xinhua Zhuang and Yun-Qing Shi for providing technical insight and constructive comments in the final editing of this paper. Our thanks to Gang Wei for allowing us to use his work on omni-directional face detection. We also thank Malcolm Slaney for pointing us to his earlier work.

9. REFERENCES

[1] M. M. Cohen and D. Massaro, "Synthesis of visible speech," Behaviour Research Methods, Instruments and Computers, Vol. 22, No. 2, pp. 260-263, April 1990.

[2] Harry McGurk and John MacDonald, "Hearing lips and seeing voices," Nature, 264: 746-748, December 1976.

[3] Michele Covell, Christoph Bregler. "Eigenpoints." *Proc. Int. Conf. Image Processing*, Lausanne, Switzerland, Vol. 3, pp. 471-474, 1996.

[4] Mingkun Li, Dongge Li, Nevenka Dimitrova, and Ishwar K. Sethi, "Audio-visual talking face detection," *Proc. International Conference on Multimedia and Expo (ICME)*, pp. 473-476, Baltimore, MD, July 2003.

[5] Malcolm Slaney and Michele Covell, "FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks," *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 814-820, November 2000.

[6] John Hershey and Javier Movellan. "Using audio-visual synchrony to locate sounds," *Proc. Advances in Neural Information Processing Systems (NIPS),* pp. 813-819, December 1999.

[7] Hani C. Yehia, Philip E. Rubin, Eric Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," Speech Communication, Vol. 26, pp. 23-43, 1998.

[8] Dongge Li, Gang Wei, Ishwar K. Sethi, N. Dimitrova, "Person Identification in TV programs," Journal on Electronic Imaging, Vol. 10, Issue. 4, pp. 930-938, October 2001.

[9] John W. Fisher III, Trevor Darrell, William T. Freeman, Paul Viola, "Learning joint statistical models for audio-visual fusion and segregation," *Advances in Neural Information Processing Systems (NIPS)*, pp. 772-778, November 2000.

[10] G. Iyengar, H. Nock, C. Neti, "Audio-visual synchrony for detection of monologues in video archives" *Proc. ICASSP*, April 2003.

[11] Ishwar K. Sethi, Ioana Coman, Brian Day, Feng Jiang, Dongge Li, Jose Segovia-Juarez, Gang Wei, and Bemon You, "Color-WISE: A system for image similarity retrieval using color," *SPIE Proc. on Storage and Retrieval for Image and Video Database VI*, vol. 3312, pp. 140-149, San Jose, CA, January 1998.

[12] M. G. Brown, J. T. Foote, G. J. Jones, K. S. Jones, and S. J. Young, "Open-vocabulary speech indexing for voice and video mail retrieval," *Proc. of ACM Multimedia 96*, pp. 307-316, Boston, MA, 1996.

[13] Fillia Makedon and Charles Owen, "Cross-modal retrieval of scripted speech audio," *SPIE Proc. On Multimedia Computing and Networking*, vol. 3310, pp. 226-235, San Jose, CA, January 1998.

[14] Gang Wei and Ishwar K. Sethi "Omni-face detection for video/image content description", *Proc. International Workshop on Multimedia Information Retrieval, in conjunction with ACM Multimedia Conference 2000, (MIR2000)*, pp. 185-189, November 2000.

[15] Wojtek Krzanowski, Principles of multivariate analysis, Oxford University Press, Oxford, 1988.

[16] Pei L. Lai and Colin Fyfe, "Canonical correlation analysis using artificial neural networks," *Proc. European Symposium on Artificial Neural Networks (ESANN)*, 1998.

[17] Barbara G. Tabachnick and Linda S. Fidell, Using multivariate statistics, Allyn and Bacon Press, 1996.

[18] Dongge Li, Ishwar K. Sethi, Nevenka Dimitrova, Tom McGee, "Classification of general audio data for content-based retrieval", Pattern Recognition Letters, Vol. 22, No. 5, pp. 533-544, April 2001.