ABSTRACT

This paper outlines several multimedia systems that utilize a multimodal approach. These systems include audiovisual based emotion recognition, image and video retrieval, and face and head tracking. Data collected from diverse sources/sensors are employed to improve the accuracy of correctly detecting, classifying, identifying, and tracking of a desired object or target. It is shown that the integration of multimodality data will be more efficient and potentially more accurate than if the data was acquired from a single source. A number of cutting-edge applications for multimodal systems will be discussed. An advanced assistance robot using the multimodal systems will be presented.

Index Terms— Data fusion, multimedia, multimodal, pattern recognition.

1. INTRODUCTION

Unimodal recognition systems usually afford low level of performance due to large intra-class variation and the noisy nature of the data. Such limitations can be partially addressed by incorporating multimodal information. The aim of multimodal media fusion is to combine multiple media sources, identify the complementary, independent, and redundant information, and generate a more complete, precise, and discriminant representation of the signals.

The fusion of multimodal information can be performed at three different levels [1]: data/feature level, score level, and decision level. Data/feature level fusion combines the original data or features extracted through certain fusion strategies. One major drawback of fusion at this level is due to the problem of “curse of dimensionality,” which is usually computationally expensive and requires a large set of training data [2]. Furthermore, the fusion problem may be difficult when different types of features are extracted from different modalities, such as the minute points for fingerprints and Principal Component Analysis (PCA) features for face images. Fusion at score level combines the scores generated from multiple classifiers using multiple modalities. To normalize all of the scores generated by different modalities in the same range, a number of different techniques have been proposed in the literature but Min-max, z-score, and \( \tanh \) normalization techniques followed by a simple sum of scores fusion method, generally outperform other techniques [3]. Decision level Fusion generates final results based on the decision from multiple modalities or classifiers using methods such as majority voting. The fusion decision level is rigid due to the limited information left. Fusion at score and decision level can be considered a special case of fusion at data/feature level [4].

The design of a multimodal media fusion system is critically dependent on the characteristics of the data as well as the requirement of the application. A multimodal biometric system can only utilize the physical and behavioral traits in isolation or a combination of both. Physical traits include face, fingerprint, iris, palm print, hand geometry and ear shape. Physical traits are considered to be unchangeable, whereas behavioral traits like hand written signatures, gait, manner of typing at keyboard, and voice are the reflection of human behavior and vary with the passage of time. A set of points that need to be considered include [4]: information sources, feature extraction method, fusion level, fusion strategy and architecture, and whether any background knowledge needs to be embedded. The information sources are determined by the application. In general, the selected modalities should be easy to collect, low cost, low noise level, fast computation, universal, diverse, and most importantly should be able to identify discriminant patterns for recognition. A multimodal emotion recognition system can utilize speech, image, video, or possibly some biological signals, like ECG, EEG.

Throughout our research, we have examined a number of publicly available multimodal databases, XM2VTS contains face and voice data [5]; MCYT contains fingerprint and handwritten signature data [6]; BIOMET contains face, voice, fingerprint, hand and handwritten signature data [7], and BioSec baseline corpus contains samples of face, voice, iris and fingerprint data [8]. These databases not only encourage researchers to work on multimodal fusion methods, but improve each modality independently.

2. EMOTION RECOGNITION

Audio and visual signals are the two major indicators of human emotion. They are easy to collect, low cost, and have
been two of the most investigated modalities for emotion recognition. Several multimodal systems have been proposed after Bolt’s pioneering system [9]. Speech and lip movements have been merged using histogram techniques [11], multivariate Gaussians [10], artificial neural networks (ANNs) [11, 12] or hidden Markov models (HMMs) [10]. In all these systems, the modalities’ probable outputs have been combined assuming conditional independence by using either the Bayes’ rule or a weighted linear combination over the mode probabilities for which the weights were adaptively determined. While time synchrony is inherently taken care of (at least partially) in the ANN-based systems described in [11, 12], this cannot be adequately addressed in the other systems. To address temporal integration of distinct modalities, a generic framework has been put forward in [13]. A popular approach is the use of a rule-based system for semantic fusion [14, 15]. Holzapfel et al. [14] have implemented a multimodal system for fusing speech and 3-D pointing gestures. They have defined a rule-based framework, using Typed Feature Structures for defining the semantic gestures. Fusion is performed based on the n-best lists generated by each of the parsers. In Mehta et al. [15], a rule based system has also been used for fusion in Human-Computer interaction. A drawback of this approach is its assumption that the individual sensors can give an unambiguous output. Gupta et al. [16] have used a multichannel parameter fusion algorithm using a weighted mean technique to classify differential brain activities. Wu et al. [17] have addressed the issue of fusion by using independent component analysis to reduce the overall dimensionality of the feature set. This is followed by a super-kernel fusion technique to fuse the individual classifier outputs. Johnston and Bangalore [18] have proposed Finite-state models for integrating multimodal input from speech and gestures. However, the model does not address ambiguity in gesture events, and can only disambiguate cues in speech modality. Kaiser et al. [19] described an approach to multimodal fusion that accounts for the uncertain nature of information sources.

In [20], we introduced a bimodal emotion recognition system using audio and visual information. A set of prosodic and phonetic features are extracted from the speech signal. Based on the speech information, a key frame is extracted from a video sequence, and a Gabor wavelet-based method is used for facial expression representation. The two streams of features are fused together at the feature level via vector concatenation. To further reduce the dimensionality of the feature space whilst finding the discriminant features for recognition, feature selection was performed. Finally, an adaptive multi-classifier scheme, which involves the analysis of individual classes and the combination of different classes is proposed for classification. A rule-based method is employed for final decision. The experimental results demonstrate that the combination of audio and visual information outperforms either of the two modalities only. Further, it is shown that some emotions are audio dominant, while others are visual dominant. The complementary relationship of these two modalities on different emotions helps to achieve higher recognition accuracy.

3. DECISION FUSION TECHNIQUE IN MULTIMEDIA DATABASES

The ever-increasing growth of multimedia information has been witnessed and experienced by human beings since the beginning of the information era. An immediate challenge resulting from the information explosion is how to intelligently manage and benefit from multimedia databases. Content-based retrieval (CBR) has been intensively studied for more than a decade, yet still remains a challenging topic [21]. In this section, we will present our recent works on fusion techniques for content-based image and video retrieval.

3.1 Fusion in Image Annotation and Retrieval

Due to the inherent ability of natural language to characterize semantics in images, much effort has been put on image annotation and retrieval jointly based on visual and textual information. In [22], both query-by-keyword and query-by-example are enabled with the measured similarity linearly combined at a subsequent step. More interestingly, researchers also endeavored to explore the integration of audio and visual information in image retrieval in mobile devices [23]. Furthermore, study on multimodal retrieval from the perspective of a user and a real application has also been conducted [24], showing that a common user can combine up to four types of modality in a search task.

While decision-level fusion puts the emphasis on the decision function, early fusion focuses on learning the representation of the joint multimodal feature space. Many models [25, 26] have been proposed which stem from latent semantic analysis and its statistical descendant, probabilistic latent semantic analysis (PLSA), originally developed for document indexing. These approaches are generally referred to as topic modeling because they assume a mixture of component distributions in the joint feature space, each of which encodes a hidden semantic topic. By introducing flexibility to the a priori distribution of the hidden topics, latent Dirichlet allocation can be considered as an extension to the PLSA. Another type of information that has proven helpful is contextual information, characterizing statistical relations across different concepts.

We recently have explored the integration of content and context with its application to image annotation and retrieval, a particular form of decision level fusion. In terms of annotation [28], the context is regarded as the
information describing the statistical inter-dependence across different concepts and the content is considered as the low-level visual features, which are combined through the Bayes’ theorem. Experimental results based on a database featuring 50 kinds of animals and 15 background concepts demonstrate the performance improvement of over those purely based on single modality, including the cross-media relevance model [30]. Different from other context-aware methods [27], we model the two modalities separately, avoiding the high computational complexity resulting from the partition function. We also attempted to deal with the semantic gap in terms of image retrieval [29], where the statistical relation across the database images is considered as the context, extracted from the past retrieval results. The context model is combined with two content-based retrieval methods, i.e. nearest-neighbor retrieval and query movement for relevance feedback (RF), and retrieval using support vector machines and active learning for RF. Experiments with 10,000 images drawn from the Corel collection showed better performance resulting from the information fusion.

3.2 Decision Fusion for Video Retrieval

Multimodal media fusion has played an important role in multimedia database and management domain. For example, they have been demonstrated to help solve challenging problems in organizing video collections, such as story segmentation, event content analysis, concept detection, retrieval and topic clustering. The problem is to integrate the different sources of evidence from many multimedia features into indexing that helps the system to effectively find what the user is seeking. In the broadcast news video domain, the text information from speech transcription and close captioning can be exploited and fused together with audio-visual features, showing promising performance [31]. In field sport video such as soccer, rugby, hockey, and Gaelic football, there are some interesting events to be analyzed and detected for video classification. An event constitutes a successful scoring incident from a sport game, e.g., a goal in soccer or hockey, a try or conversion in rugby, a point in Gaelic football, etc. Each one of these event can be uniquely characterized by features such as “Crowd Image”, “Speech-Band Audio Activity”, and “On-Screen Graphics”. These are detected by event-defining feature detectors, each of which provides evidence which is combined by support vector machine (SVM) to infer the occurrence of an event [32].

The techniques discussed above are some examples of multimodal media fusion techniques in the following two categories: data fusion and decision fusion. In the following discussion, we focused mainly on the score fusion technique using two-layer architecture. The scores are independently obtained, which are then combined. The lower layer contains local experts, each of which produces a local score based on a single modality. The upper layer combines the score. In [33], we proposed a nonlinear score fusion using SVM for classification of movie clips. An adaptive video indexing for visual feature and Laplacian mixture model for audio feature, are fused at the late fusion stage. This is input to SVM to learn semantic concepts, such as Love scene, Fighting, and Ship crashing.

4. FACE AND HEAD TRACKING

Face detection and head tracking are considered to be essential requirements for intelligent vision-based human computer interaction systems, such as video surveillance, face recognition, emotion recognition and face database management. An effective face detection system can detect the presence of all human faces in an image and give rough estimations of the positions of all detected faces in real time. Many face detection techniques have been studied extensively in the past decades, including feature based methods, using geometric information such as skin color, geometric shapes, motion information, and machine-learning based approaches like neural networks, Gaussian mixtures, support vector machines and statistical modeling [34-36]. In [37], we have introduced a fast and robust face detection method which incorporates local normalization with optimal adaptive correlation (OAC) technique into a conventional face detector to alleviate illumination variation problem. The experiment results demonstrate that the method improved the face detection rate and reduced the processing time at the same time. In our recent automatic fiducial point detection system, fiducial points are detected in video sequences using scale invariant feature based Adaboost classifiers for facial expressions analysis. The results demonstrate our method achieves a good performance.

5. APPLICATION IN ROBOTICS

With Multimodal systems allowing for the combination of many recognition algorithms, we have had success combining many recognition systems together for a robot capable of mobile security patrol, domestic chores (e.g., for the elderly and physically limited), and for medical assistance [38]. The robot has an onboard computer responsible for system operation and diagnosing, but the more complicated recognition algorithms are performed remotely for face, emotion, hand gesture, body movement, and speech recognition. The system is also capable of self navigation, self-docking (for charging) and identifying its location through constellation navigation based on markers in the corners of ceilings. The robot can roam and perform mundane human tasks and alert human users (who can access what the robot is experiencing remotely) when more complex choices are required. We are in the process of porting over these system designs from the robot to virtual reality applications. This creates the possibility for
individuals, such as engineers, emergency specialists, or doctors to provide remote assistance in situations where they could not be physically present. These systems create a stronger bridge between people and computers, but also between people communicating between computers.

6. CONCLUSIONS

In this paper, we introduced and discussed important issues in the research, design and applications of multimodal systems. There was a time when multimodal processing was too complex for both processing and concept. However, with the advances in computing technology, as well as the current broadening of multimedia research and disciplines, it is anticipated that we will see much more intuitive and dynamic multimodal systems that is increasingly similar to the learning and functioning of the human brain.

REFERENCES


