Multimodal Data: Acquisition, Processing, Storage and Exploration

ABSTRACT

This paper is an introduction to the not-so-young but now rediscovered world of multimodal data. Within the context of signal processing, a lot of papers have recently appeared showing new algorithms to deal with simultaneous data from multiple sources. This has been motivated by what is known as the Big Data phenomenon: cheap acquisition and cheap storage of extremely large collections of data. When dealing with this kind of data, new issues appear but new things are made possible. The main aspects of this summary article address this and may give answers to the newcomers on the field. The first issue is very physical: how to record multimodal data. After this acquisition step, it is crucial to guarantee that the streams are synchronized. This problem, as well as some others, can be addressed with particular signal processing techniques. The next issue is how to store the files in a structured way that allows finding them later, and finally how to visualize the streams. The implications for these techniques are broad and differ a lot; some particular cases are presented illustrating how multimodal data can be exploited. For instance, combining information from multiple biosensors provides doctors a better understanding of some injuries, and by joining smartphone video streams one can recreate a concert to an extent never seen before.

KEYWORDS

multimodal, signal processing, big data, data storage, data visualization

1 Introduction

Signal processing is a mature field, and researchers have developed algorithms and techniques to carefully analyze data far beyond the limits of human perception. However, there is always a big constraint: the data itself. Our world is so complex that the only way to correctly model is through uncertainty, and this means sacrificing details for the sake of simplicity. Most of those algorithms aim to reduce this uncertainty in order to discover the latent, inherent - some might say "pure" - data hidden behind a wall of noise and randomness.

However, very often the uncertainty is so high that we say that the underlying data is partially observable; that is, no matter how dimensions the input has, we can get its inner structure only to a certain extent. For instance, video signals can be interpreted as high dimensional vectors that change over time, but if an object occludes (stays in front of) what we are analyzing, we have no chance of seeing it. If our goal is not the data itself but a higher level representation of it, a way to overcome this problem is through the use of multimodality: multiple, independent data sources that tell us different aspects of reality we can combine afterwards to get a more accurate knowledge of it. In the case of video occlusion, if we wanted to recognize the a subject we could place a microphone in the room as well in order to recognize their voice among a speaker database, and then combine the visual and auditory clues. One notorious application of this kind of techniques is found in medicine, where doctors can get both the temporal resolution of EEG and the spatial resolution of fMRI at the same time to predict what happens inside the brain with much more precision (Biessmann, 2011).

Another viewpoint of multimodal data is interdependence analysis; that is, how each source relates to the others within a common context. For instance, in a conversational environment one might wonder what is the relation between what people say and their movements. This approach requires a more concise definition of multimodality, which I will describe below.

This paper is organized as follows: Section 2 discusses the data dimensionality in multimodal environments. Section 3 describes the main issues in data acquisition and synchronization, as well as some approaches to solve them. Section 4 addresses the problem of storage, edition and visualization of multimodal data. Finally, Section 5 summarizes all the above and gives some examples of real-world multimodal data analysis scenarios.

2 Data dimensionality

In the context of modality, the term source can easily lead to confusion. It might seem reasonable to assume that in multimodal environments each multidimensional data signal represents a modality; that is, each signal provides independent information about different aspects of reality. For instance, in traditional multimedia experiments, image gives visual clues about shapes or motion and sound gives auditory clues about events such as impacts. Then, source would refer both to the device that generates the signal and the features or modes it represents

This assumption is not always valid, and depending on the case it must be refined - or relaxed. It might be the case that multimodality can be omitted entirely by concatenating all signals (i.e. two tridimensional signals would lead to a single 6-dimensional signal) and considering all of them a single feature. This approach is valid for some classification problems where the classifier is smart enough to take only the relevant aspects from each dimension.

Most of the times, especially when dealing with correlated sensors, each signal or event each dimension of the signal is a mixture of different features. Then, it is very important to distinguish between modes or features and signals or sources. It is very different to deal with a lot of signals that convey information about a few features or with a few signals from which a lot of features have to be extracted, and completely different approaches are required. Section 5 shows some techniques to perform these separations.

Apart from the number of modalities, it is important to think about other dimensions of the data. If the experiment focuses on a single recording of all modalities it will be very important to ensure that we have the maximum precision and synchronization in the measures. However, if the experiment focuses on recording a lot of times and the comparing the outcomes a tradeoff between accuracy and volume of data must be found to prevent the system from overloading. The recent advances in data storage and capabilities have relaxed this condition. It is now possible to record a lot of streams for a long time with a minimal load of the resources.

3 Data acquisition and synchronization

The most important thing of multimodal analysis is, of course, the data itself. Therefore, it is crucial to take care when capturing it in order to minimize uncertainty in terms of noise. However, there is one aspect that is sometimes left behind and in some cases it the most relevant: synchronization between data streams. In order to compare or fuse the features from each modality, the timing between them must be the same as in the recording environment. If the features are spread among the signals, there are techniques that allow the synchronization *a posteriori* of the signals that will be described later. However, if the sources are highly independent or their temporal relationship is the object of study the timing must be fixed beforehand.

Many protocols exist that allow embedding time codes into the data. By using these techniques, markers with a universal clock time are put into the data at regular times. The SMPTE time code is a very popular format in the audiovisual field, but there are others. By comparing the time stamps of the different files, it is very easy to align them even if their recording started at different times. Very often, researchers build their own acquisition tools in order to manage all this processes efficiently (Cappelletti, 2008).

It is possible to align data streams afterwards without clues by finding the time offset between them. In order to accomplish this, it is necessary that the streams have redundancy; that is, that some features are represented in more than one stream. Then, by computing the cross-correlation between each pair of streams the best match can be found; it will be the offset that makes the signals more similar, and this will mean that the common features coincide. However, this method requires a definition of similarity between data streams, which is not trivial if they come from sources of different nature such as audio or video.

If the researcher is comparing different recordings of the same experiment, there are techniques that allow the alignment of them taking into account that they are not simultaneous. This requires a non-linear mapping of the time codes of the signals, because in one experiment one part could be slower and another one faster than in another take. If the streams are low-level data, Dynamic Time Warping is often used. A modification of the algorithm allows the alignment of multimodal data (Wöllmer, 2009). If the streams are high-level data, as happens with semantic information, higher-level strategies such as (Hidden) Markov Models can be used.

Some multimodal data frameworks provide data acquisition and synchronization tools that ease all this processes a lot. The Observer XT (Zimmerman, 2009) and the EyesWeb XMI (Camurri, 2007) are some examples.

4 Storage, edition and visualization

When the volume of the recorded data is high or there are many people involved in the project, it is very important to carefully choose or design the storage system where the data will he held. Since recording multimodal data is often very expensive, the data is often put online so that other people can benefit from it or to allow the partners to explore it remotely. These databases are called repositories, and some of them even provide an interface that allows data visualization directly on the web. The RepoVizz (Mayor, 2011) and ANNEX (Berck, 2006) are some examples.

In most of the cases, the raw data will not be the most relevant aspect but the annotations experts on the field make on top of it. These annotations can refer to things that happen in a particular time instant (events), or to what happens during a period of time. If many people are annotating the data at the same time, a distributed system is required. MySQL databases have proved to be a good choice (Kipp, 2012).

Apart from annotations, it is common to pre-process the data before analyzing the features in order to isolate them. The input data can have many sources and each source can have many dimensions, and there are many techniques that try to reduce this redundancy down to the single features, such as Principal Component Analysis or Independent Component Analysis (Wu, 2004).

5 Conclusions

Many peculiarities of multimodal analysis have been presented, and a lot of examples have been mentioned in the way. In short, multimodal analysis is about recording different aspects of the reality in order to have more information about the events we want to study, which is called multimodal fusion (Atrey, 2010), or in order to study the relationship between these aspects (interdependence analysis). It is very important to isolate these aspects or features from the multiple sources we have, and there are many techniques for doing so. It might be useful to put the data in an online repository in order to work collaboratively, but in any case the most important thing is to ensure that the different sources (and features) are time-aligned. How to perform further analysis of the data is beyond the scope of this paper, but some information can be found in the references.

There are many examples of multimodal analysis in the literature. In the context of music analysis, some studies have been carried that analyze how musicians perform in terms of rhythm or intonation when they play in group (Papiotis, 2012), using motion capture data and audio streams. This is an example of interdependence analysis. In the context of data mining, a project by the Florida International University aimed to detect soccer goals by analyzing the video feed of the cameras and the sound made by the ball (Chen, 2004).

References

(Zimmerman, 2009) P. H. Zimmerman, J. Elizabeth Bolhuis et al. (2009) The Observer XT: A tool for the integration and synchronization of multimodal signals. Behavior Research Methods n.41 (3), pp. 731-735, Springer-Verlag.

(Cappelletti, 2008) A. Cappelletti et al. (2008) A multimodal data collection of daily activiites in a real instrumented apartment. Proceedings of the Workshops of the 6th International Conference for Language Resources and Evaluation (LREC-08), pp. 20-26m Marakech, Morocco.

(Papiotis, 2012) P. Papiotis, M. Marchini, E. Maestre (2012) Computational Analysis of Solo Versus Ensemble Performance in String Quartets: Intonation and Dynamics. Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC12), Thessaliniki, Greece.

(Chen, 2004) Shu-Ching Chen et al. (2004) A Decision Tree-based Multimodal Data Mining Framework for Soccer Goal Detection. Proceedings of the IEEE International Conference on Multimedia and Expo (ICME04), pp. 265-268.

(Wöllmer, 2009). M. Wöllmer et al. (2009) A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams. Neurocomputing n.73, pp. 366-380.

(Atrey, 2010). P. K. Atrey et al. (2010) Multimodal fusion for multimedia analysis: a survey. Multimedia Systems n.16, pp. 345-379.

(Camurri, 2007) A. Camurri, et al. (2007) A Platform for Real-Time Multimodal Processing. Proceedings of the 4th Sound and Music Computing Conference (SMC'07), pp.354-358. Lefkada, Greece.

(Wu, 2004) Y. Wu, et al. (2004) Optimal Multimodal Fusion for Multimedia Data Analysis. Proceedings of the 12th annual ACM international conference on Multimedia (MULTIEMDIA'04), pp. 572-579. New York, USA.

(Biessmann, 2011) F. Biessmann, et al. (2011) Analysis of Multimodal Neuroimaging Data. IEEE Reviews in biomedical engineering n.4, pp. 26-58.

(Kipp, 2012) M. Kipp (2012) Multimedia Annotation, Querying and Analysis in ANVIL. Multimedia Information Extraction: Advances in Video, Audio, and Imagery Analysis for Search, Data Mining, Surveillance, and Authoring (ed M. T. Maybury), John Wiley & Sons, Inc., Ch. 21. Hoboken, USA.

(Mayor, 2011) O. Mayor, J. Llop, E. Maestre (2011) Repovizz: a multimodal on-line database and browsing tool for music performance research. Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011). Miami, USA.

(Berck, 2006) P. Berck, A. Russel. (2006) ANNEX – a web-based Framework for Exploiting Annotated Media Resources. Proceedings of the 5th International Conference for Language Resources and Evaluation (LREC-06). Genoa, Italy.