Neurocomputing 144 (2014) 128–137

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Clustering and retrieval of video shots based on natural stimulus fMRI

Junwei Han [a], Xiang Ji [a], Xintao Hu [a], Jungong Han [b,*], Tianming Liu [c]

[a] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China
[b] Civolution Technology, Eindhoven, The Netherlands
[c] Department of Computer Science and Bioimaging Research Center, The University of Georgia, Boyd 420, Athens, USA

ARTICLE INFO

Article history:
Received 29 June 2013
Received in revised form
20 October 2013
Accepted 6 November 2013
Available online 12 June 2014

Keywords:
Video clustering
Video retrieval
Functional magnetic resonance imaging
Feature integration

ABSTRACT

Functional magnetic resonance imaging (fMRI) is a powerful tool to probe the human brain's perception and cognition. Besides being extensively exploited in the clinical applications, fMRI technique is also useful to human's ordinary life. In this paper, we investigate a novel application of leveraging fMRI techniques to video clustering and retrieval. In the proposed work, we successfully integrate semantic human-centric features derived from natural stimulus fMRI data and low-level visual-audio features to facilitate video clustering and retrieval, which is a significant innovation compared to the previous works relying on either fMRI-derived features or low-level visual-audio features. Our system consists of several algorithmic modules. First, fMRI data when the subjects are watching video shot samples are acquired. Then a newly developed brain networks localization system is employed to locate the cortical regions of interests (ROIs) for each individual subject. The functional interactions computed by wavelet transform coherence are quantified, from which the human-centric features are derived. Afterwards, the Gaussian process regression model mapping visual-audio feature space to an fMRI-derived feature space is trained, given the training samples. The trained model is then adopted to predict fMRI-derived features for videos without the fMRI data. Finally, the multi-modal spectral clustering and multi-modal ranking algorithm are adopted and proposed to integrate these two heterogeneous features for video clustering and retrieval, respectively. Our experiment on TRECVID database has demonstrated the precision of video clustering and retrieval can be substantially improved by integration of visual-audio features and fMRI-derived features.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Functional magnetic resonance imaging (fMRI) is a neuroimaging method that examines brain function and identifies areas of increased or decreased neuronal activity based on the changes in blood oxygen level-dependent (BOLD) signal [33]. As a non-invasive technique, fMRI enables to localize brain activations in a single subject and recognize abnormalities in these activation patterns, which may exist in populations of patients. Therefore, in the past decade, fMRI techniques have been widely applied to clinical applications such as diagnosis of disease [34] and pre-surgical planning [12].

In the past five years, we have observed an emerging direction [4–9,13] in how the human brain responds to natural stimulus such as image/video watching in neuroimaging and medical image field. Natural stimulus fMRI during a movie watching enables us to look into the dynamics of how video stream regulates the brain's perception and cognition. Its advantage is that the human subjects are naturally engaged in the perception and cognition of the multimedia streams. Despite higher complexity comparing with traditional task-based and resting state fMRI, natural stimulus fMRI is more in an uncontrolled natural environment, which is well suited to study the functional interactions and dynamics among brain regions in response to multimedia stimuli.

In image and video understanding field, content-based image/video clustering and retrieval has been extensively studied. These clustering and retrieval methods [1–3] have demonstrated good performances due to their ability of capturing underlying geometrical structure of the database. However, these methods mainly adopt low-level visual features such as color, texture, shape, and motion to measure the similarity between images/videos. Their results are still unsatisfactory because of the well-known semantic gap. A growing body of research has realized this problem and suggested integrating semantic features to enhance the performance.

The natural stimulus fMRI technique may render a new view of understanding images/videos semantically via monitoring human brain activity. This new research stream is based on the premise that human brains are the end users and evaluators of multimedia content and representation, and quantitative modeling of

* Corresponding author.
E-mail address: jungonghan77@gmail.com (J. Han).

http://dx.doi.org/10.1016/j.neucom.2013.11.052
0925-2312/© 2014 Elsevier B.V. All rights reserved.

the dynamics and interactions among multimedia streams and brain's responses can provide meaningful guidelines for multimedia understanding. Lately, a few literatures have been published along this research direction. In [4], Walther et al. conducted pattern analysis based on fMRI data to examine which brain regions can differentiate natural scene categories. In [5,6], Hu et al. proposed a video classification system by correlating fMRI-measured brain responses and audio-visual features by using the *Principal Component Analysis-Canonical-Correlation Analysis* (PCA-CCA) algorithm. Recently, as shown in [7–9], instead of low-level visual-audio features, fMRI-derived features were utilized for video retrieval and audio classification, which indicated the capability of fMRI-derived features is significantly better than that of low-level features. However, it is logical to expect that the integration of these two types of features can lead to an even better performance. This problem has not been thoroughly investigated yet, as far as we know.

The flowchart of the proposed method is shown in Fig. 1. Let us briefly go through each building block. First, a number of video shots are randomly selected as training samples. Then, natural stimulus fMRI data are acquired when subjects are watching the training samples. The brain networks of each subject are localized using the DICCCOL system [10]. Second, the brain's responses among the identified brain networks are quantified as the connectivity matrix of wavelet transform coherence (WTC) [14] using the fMRI data, which outperforms *Pearson* correlation coefficients (PCC) used in [5–9]. Afterwards, the relief-F [15] and correlation-based feature selection (CFS) [16] are performed to select the discriminative and representative elements from the connectivity matrices, resulting in the fMRI-derived features. Third, due to the high cost of fMRI scanning, we train an fMRI-derived feature prediction model using low-level visual-audio features via a Gaussian process regression (GPR) model. The trained model is then used to predict fMRI-derived features for those video shots without having fMRI data. Finally, the *multi-modal spectral clustering* (MMSC) and a novel *multi-modal ranking* (MMR) algorithm are utilized to integrate fMRI-derived features and low-level visual-audio features for the task of video shot clustering and retrieval.

The main contributions of this paper can be summarized as follows. First, unlike the previous works [5–9] that highly rely on the fMRI-derived features, our work integrates the strengths of both fMRI-derived features and low-level visual-audio features for video clustering and retrieving. Our motivation is that the fMRI-derived features and low-level visual-audio features have their own strengths and can be complementary to each other. Fig. 2 shows a positive exemplar video shot for fMRI-derived features, which has higher intra-class similarity and lower inter-class similarity in the fMRI-derived features space compared with that in the low-level visual-audio feature space. In contrast, Fig. 3 shows a positive exemplar video shot for low-level visual-audio

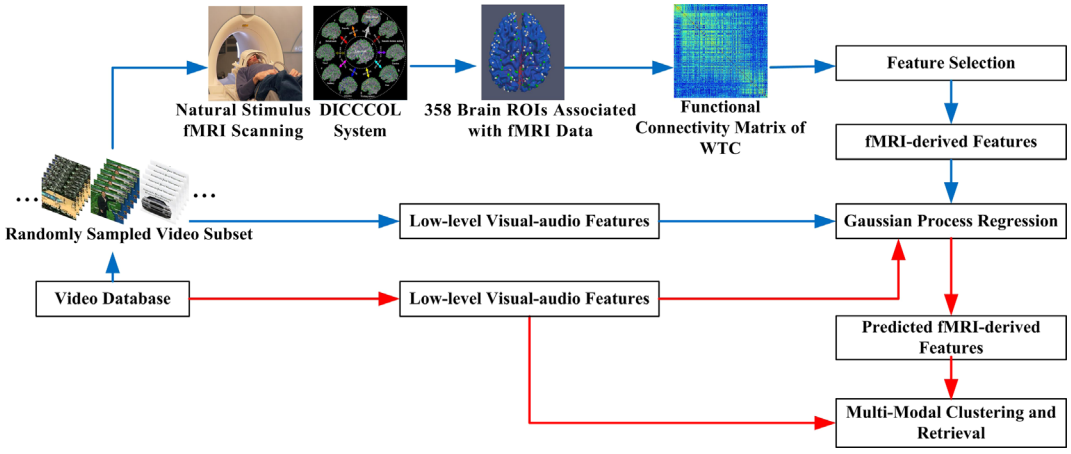

Fig. 1. The flowchart of the proposed method. Procedures indicated by blue arrows form the training stage and procedures indicated by red arrows form the testing stage. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

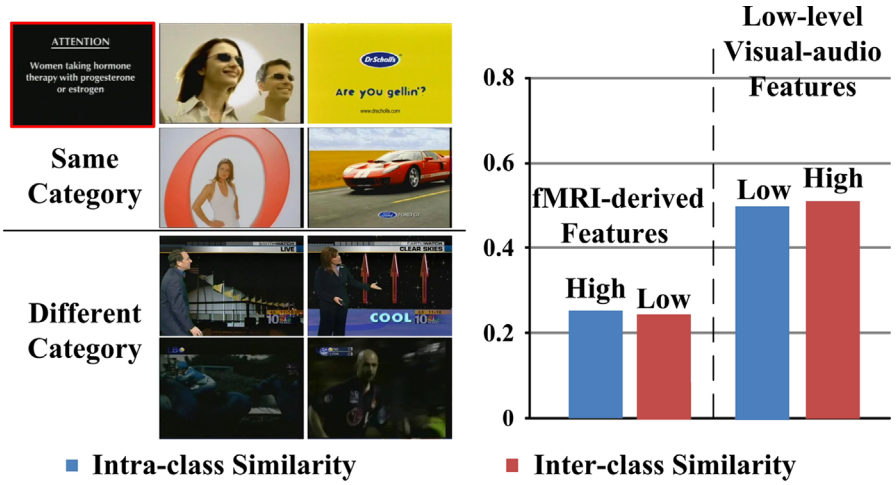

Fig. 2. A commercial video shot that has higher intra-class similarity and lower inter-class similarity in fMRI-derived feature space compared with that in low-level visual-audio feature space.

Download English Version:

https://daneshyari.com/en/article/412188

Download Persian Version:

https://daneshyari.com/article/412188

Daneshyari.com