Multimodal Sentiment Analysis Using Ensemble Classifiers

Author One
author1@illinois.edu

Esteban Gomez
gomezsa2@illinois.edu

Author Two
author2@illinois.edu

ABSTRACT

Sentiment analysis is the classification problem where an algorithm recognizes the primary emotions that a media input evokes. Emotions are subjective, so classifying media into distinct groups is a challenging problem. Most human subjects agree in broad strokes on emotional classifications. However it is not uncommon to find media inputs where there is no consensus, leading to inconsistencies in class groupings. As a result, sentiment analysis is more challenging to incorporate into real-world applications because an algorithms recognition accuracy might differ between classes.

This paper introduces a multimodal approach to the sentiment analysis problem. This method leverages the variety of information available to minimize class inconsistencies. By analyzing songs through two independent media spaces — audio and lyrics — the algorithm combines the inherent emotional information to reach a single overall classification.

Keywords: Signal Processing, Sentiment Analysis, Machine Learning, Feature Fusion, Multimodal Classification

1. INTRODUCTION

Electroencephalography (EEG) is a common method of reading electrical activity in the brain for research and medical diagnosis.

1.1 Motivation

Signals are notoriously difficult to manipulate, detect, and correct. Particularly, electroencephalography (EEG) signals are often very convoluted with muscle movement (5 - 200 hz) and power grid pollution (around 60 hz). The method described is intended to remove as much signal pollution as possible related to muscle movement using a standard laptop camera or web-cam.

The methods currently used to correct or remove noise from muscle movement in EEG signals require the manual training of a classifier every time a subject wears the headset. Every time a headset is placed on the head the connection to the scalp is altered slightly, and any classifier trained during a previous session can have issues identifying noise from useful signals.

The method described is designed to remove the need to train any classifier by finding correlations between movement and the ICA components of an EEG signal. If there is a correlation this can be used to identify the specific component and either correct it or mark it as a possible error. The advantage of this is that there is no need to train a new classifier every time an EEG is used, which takes considerable time, and noise can be more accurately identified.

Research and neuro-feedback

The research in the field of neuro-feedback is still in its infancy, but there has been some strides in making it useful for consumers. One particularly promising area of research is that of steady state visually evolved potentials or (SSVEP), which can be used as a brain-computer interface (BCI). An application of this would be the chess game created in Dr. Bretl's lab at the University of Illinois[**?**], which enabled users to play a game of chess using a robotic arm and flashing lights.

SSVEP works by stimulating the brain with a flashing light at a particular frequency, say 7 hz. This creates action potentials (neural firing) at the same frequency, and these signals can then be picked up on an EEG with a very high accuracy (85 - 90 percent). Unfortunately, SSVEP responses only seem to work (best) between the frequency of 5 hz and 30 hz, the range of frequencies heavily effected by muscle movement. Thus, if the noise due to muscle movement can be reduced then the accuracy of SSVEP detection, one of the most promising methods used in BCI today, can be improved as well.

Consumer headsets and untrained users

Most of the brain-computer interfaces today require a medical grade EEG. However, with the introduction of Emotiv and OpenBCI headsets, there is now a surplus of consumer grade EEG headsets. The downside to this is that these EEGs are used in uncontrolled environments without expert handling. Requiring a whole new range of EEG signal analysis or collection needs to be developed, in order to account for the increase in noise, else many of these EEGs will be next to useless.

This method can account for some noise and correct the signals, without constant manual training of a classifier, and without needing any input from the subject. This means that although the subject may not be an expert, the EEG signal can still be corrected/cleaned, and the consumer EEG can be used to a greater effect. The requirement of training a classifier or reviewing the EEG data is removed, hence it is more consumer friendly, and less error prone.

2. GENERATING A MOTION SIGNAL

The EEG output is a mixture of the signals resulting from many simultaneous brain processes. Since at any moment the brain is processing and responding to many stimuli, it is expected to see a very complex response from the EEG. Outside noise that is reflected in the signals include motion artifacts from muscle twitches and other type of movements. Since these responses are of little significance, all effects of movement are categorized as noise. By combining video data with the EEG, it becomes possible to make a more accurate distinction between the undesired and desired components of the brain signals. To provide an insightful analysis of the relation between the video and the EEG output, it is necessary to convert the video into a facial motion signal. An additional complication is that the location of the EEG sensor determines how strong a motion artifact will be recorded, which means that the motion signal must have a spatial component alongside its intensity component.

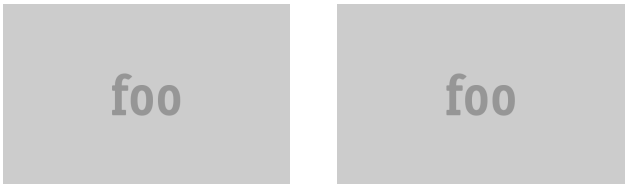
2.1 Facial detection and image stabilization

For the motion signal to contain useful information, each area of the face must be compared to the same area across multiple frames. If the user changes position without making any facial gestures, the signal should be able to reflect low motion while tracking the user in the frame. As a result, the algorithm must be shift-invariant for the face location but very sensitive to changes within the region of interest (ROI).

The region of interest was extracted using OpenCV's pretrained Haar-Cascade[**?**] facial templates. This returns a region that contains the user's face determined by its coordinates x. However, there is no guarantee that the coordinates of the ROI will be exact between one frame and the next, even if the user did not move. There is always coordinate variation that shifts the ROI slightly in any direction. The unwanted jitter in the ROI leads to adding extra noise to the motion signal, and so it is possible to interpret x as a combination of the true location $\hat{x}$, and random noise x_{noise}. Before extracting the face image from the overall frame, it is necessary to approximate the ROI's coordinates such that the random noise is minimized.

$$x = \hat{x} + x_{noise}$$

Under the assumption that x is a random variable, it is possible to minimize noise by making multiple measurements to decrease its variance. As a result, $\hat{x}$ was approximated by averaging the output of two different facial detection calls each with its own facial template. Although this average increases confidence of the location of the face in the frame, it does very little to handle inter-frame jitter and displacement. To account for the necessary shift-invariance, a ten frame buffer was added that kept track of the last nine $\hat{x}$ values, so that the ROI was defined by the average of the current location and that of the previous nine frames. As a result the ROI smoothly follows the user's face throughout the frame, while still being sensitive to what occurs inside. In the case where both Haar-Cascade[**?**] calls fail, the ROI extracted is based on the average $\hat{x}$ from the buffer. Since the user will likely move very little within ten frames, this approach permits using recent history to predict where the face is located.



(a) Face image before blink (b) Face image during blink

Figure 1: Image processing for a blink

2.2 Motion extraction

After a shift invariant ROI has been extracted it becomes considerably simpler to track changes in the face image, mainly because the different areas of the face now line up. Any motion detection scale will measure the overall changes throughout the entire face. However, as mentioned earlier, to perform useful analysis of the EEG signal it is necessary to also obtain some spatial information. As a result, a 4 x 4 grid is overlaid on the ROI to divide it into sixteen different regions. By breaking up the image in this fashion, instead of getting a single value that conveys all changes, there are sixteen different values that contain the motion of each portion of the face. On Figure 1, the overlaid grid that divides the image concentrates most of the movement on the two middle quadrants of the second row.

The grid images provide greater precision for interpreting the motion data, which enables a more detailed analysis against the EEG signal. The grid image is then converted into a series of delta images[**?**], that contain high values in regions where there was motion and low values that remained constant. This algorithm consists in measuring the pixel changes between the current, the previous, and the next frame. The delta images in Figure 2 correspond to the grid images previously shown. It is clear that before the blink, there was little motion while there is a concentration of high values around the eyes during the blink. The pixel values for each quadrant of the delta image are summed up to get an overall value representing the amount of motion for that quadrant in that frame.

2.3 Results

This technique for converting a video into a motion signal seems to provide intuitive and representative results. Since the signal conveys both intensity and location information, it is easy to determine the moments where different events occurred in the video. Figure 3a shows the change of all

(a) Delta image before blink

(b) Delta image during blink

Figure 2: Delta images during a blink

sixteen signals in time, arranged so that they correspond to the respective location on the subject's face. Barring the training period where the buffer was getting filled during the beginning, the only high-valued features align with blinks. Notice that all regions except 6 and 7 remain relatively quiet during the session. Figure 3b shows a close up of the signals around the blink that was presented in Figures 1 and 2. Note that the region showed in Figure 3b is the portion between the red lines in Figure 3a. The red line in Figure 3b shows the moment where the presented blink occurred. Like seen in Figure 2b, all the regions are quiet except for the middle two on the second row.

3. GATHERING EEG DATA

One of the goals of this research was to provide a solution to remove artifacts from commodity EEG headsets, which are of much lower quality than those found in the medical field. A low-cost headset that provides 14 channels, the Emotiv Epoc, was chosen. It is known to not perform as well as medical grade devices[**?**]. Next, a experiment was developed to collect the readings off of the 14 electrodes while simultaneously capturing video from a laptop web-cam.

3.1 Experimental setup

The experiment consisted of the subjects performing various facial expressions or motions while looking at the camera. Each action was performed for 20 seconds and the subjects would switch actions when the screen indicated. The last action in the sequence was to look at a screen to the subject's right which was flashing a box. This was meant to be used to establish a baseline reading and examine SSVEP (see Section 1.1), although that proved difficult.

Some example expressions were scratching the face, tugging ears, pretending to talk, blinking, and scratching the top of the head.

Each test took approximately 5 minutes in total.

3.2 Data processing and cleaning

Using start points on the video and EEG data, the two signals were aligned by hand and cropped to be within half of a second of each other. Some of the video data was hard to process due to various factors such as skin color and large amounts of hair preventing facial detection. Additionally, some subjects exaggerated movements greatly, which made it difficult to track the face. However, most of these obstacles were overcome using the motion signal construction technique in the previous section.

(a) Motion signals for recording

(b) Motion signals around the blink

Figure 3: Motion response to a blink

The EEG data, E, then contained 14 channels at a sampling rate of 129 Hz and the motion data, M, consisted of 16 channels at a sampling rate of 15 Hz. Both of these datasets were normalized to unit variance and zero mean before continuing. The motion data and EEG data were re-sampled to 516 Hz, which is the standard sampling rate in many of the medical grade headsets.

4. REMOVING MOTION COMPONENTS

The method for removing ICA components that contain motion involves first segmenting the data into small windows, E_w and M_w. The authors found success with windows of 0.25 seconds, although they experienced acceptable results up to 3 second windows. Different segment sizes are appropriate for removing different motions. For example, a blink occurs in less than 0.5 seconds, but scratching the head or face might take longer.

Assuming EEG artifacts due to motion and normal EEG data would be statistically independent, blind source separation should be possible by performing independent component analysis (ICA) on E_w to extract the components. Once found, "bad" components could be removed and the original signal could be reconstructed from the remaining ones.

Psuedocode for this procedure can be found in Algorithm 1.

4.1 Similarity measure for motion and EEG signal

A measure for deciding which components to remove was needed. Some similar methods use a classifier that is trained by EEG experts on artifacts in the subject's data for that particular session. This is problematic as it requires experts to ensure that the headset was applied to the subject in a particular way to ensure quality contacts with the skin, that the test was controlled for movement, and to later clean out artifacts. To achieve the same results on a commodity EEG in the home without an expert present, an unsupervised metric was needed. A simple normed cross-correlation (`xcorr`) metric, c was chosen. This allowed for slight shift invariance in the two signals, as they may not be perfectly aligned. However, due to segmenting the invariance an not extend beyond the window's edges. The metric shown below.

$$c(C, M_w) = \frac{\texttt{max}\ |\texttt{xcorr}(C, M_w)|}{||C|| * ||M_w||}$$

This metric was indicator of how likely an EEG segment component correlated with a motion segment, with a value of 1 being exactly correlated and 0 being no correlation with any lag value. After experimentation with some various thresholds, empirical results indicated that 0.8 removed many ICA components that contributed to artifacts near high levels of motion in the time series. The original signal was then reconstructed without the bad components.

4.2 Note on ICA algorithms

While several algorithms to choose components were tested, the one that performed best was the ICA implementation in EEGLab[**?**]. FastICA[**?**] performed well in some segments of the data, but it would fail to converge in others.

Algorithm 1 Remove undesired ICA components from EEG signal

```
1: procedure CleanSignal(E, M, step, threshold)
2:     re-sample E and M to 516 Hz if needed
3:     for each E_w, M_w segment of length step do
4:         Bad ← {}
5:         ICAComponents, W ← ICA(E_w)
6:         for comp ∈ ICAComponents do
7:             if c(comp, M_w) > threshold then
8:                 Bad ← Bad ⋃ {comp}
9:         Activations ← W · E_w
10:        Good ← ICAComponents \ Bad
11:        E'_w ← W^{-1}(:, Good) · Activations(Good, :)
12:        append E'_w to reconstruction
```

5. RESULTS

Overall, results of this method are rather mixed. While the method presented did identify and remove some motion-related components, it was not nearly as effective as desired. Typically, only noise with high amplitude that coincided with motion of high amplitude was removed. Additionally, analysis in the frequency domain indicates that some artifacts are introduced by the reconstruction that may harm analysis of the underlying EEG signal.

Attempts to reduce the undesired component threshold (0.8) to remove more noise resulted in large chunks of valid data being thrown out.

5.1 Reconstruction

The reconstruction, for the most part, was close to the original signal. Some large portions of noise that correlated well were removed. Figure 4a highlights a segment of the signal in which some components were removed due to high correlation with the motion signal while the subject was talking.

In contrast, Figure 4b depicts two segments where no components were removed due to low correlation. The left segment has noise in the EEG signal but no movement (the subject's movement was not in range of the camera). The right segment has some motion but little correlated noise.

In both cases, relevant EEG data seems to have been preserved. While no re-synthesis from a component analysis can provide a perfect reconstruction, this reconstruction is close.

5.2 Power Spectral Density

Power Spectral Density (PSD) is a well known method of analysis for EEG signals[**?**]. The PSD shows the amount of different frequencies present in the signal. Muscle spasms and eye twitches should appear in the 5-200 Hz range. Therefore, if the method is successful then the cleaned data should have a similar PSD to the original data, perhaps with lees frequencies in those ranges. However, comparisons of the power of the boxed ranges in Figure 5a to Figure 5b reveals that the power in these areas *increased*.

While more analysis must be done before saying the definitive cause of these artifacts, it is possible that the small segmentation windows cause frequencies to be introduced

(a) PSD of original EEG signal

(b) PSD of cleaned signal. Note the amplitudes in the areas highlighted by the boxes.

Figure 5: Analysis of the frequency domain before and after cleaning.

(a) High correlation

(b) Low correlation

Figure 4: In (a), the artifact is detected due to high correlation with motion. Since the noise in (b) was unrelated to subject motion in the video feed, it was not removed. In both images, the signals presented are (in order): EEG signal, motion signal, removed components, and reconstructed signal.

when components are removed. The sliding filter discussed in Section 6.2 may alleviate this issue.

6. FURTHER WORK

The authors recognize that the results are unsatisfying and could be vastly improved. Many avenues for enhancement have been identified, but they unfortunately could not be implemented due to time constraints.

6.1 Success measure

The analysis of the results is done by visual inspection of the time and frequency domains without an obvious success metric. Typical methods to remove artifacts with blind source separation have experts code the artifacts by spatial location using heuristics[**?**, **?**]. Construction of a comprehensive data-set with annotated artifacts and associated video would be fantastic and allow for comparison against other method on the same subjects. It would also allow for this method to be evaluated objectively by ratio of artifacts removed and ratio of desired signal preserved.

6.2 Sliding filter to replace segmentation

Segmentation into discrete chunks can lead to alignment issues between the motion and EEG signals or make artifact features unrecognizable. Allows for shift invariance across the entire signal would be ideal.

Additionally, segmentation is believed to be the cause of the artifacts introduced in the power spectral density of the reconstruction. Hopefully a sliding filter with overlapping windows would fix this.

6.3 Other algorithms for source separation

While FastICA[**?**] was compared to another implementation of ICA in EEGLab[**?**], few other source separation techniques were attempted. Analysis of wavelet decomposition and non-negative matrix factoring as possible algorithms would be interesting.

Cross-correlation was also the only similarity measure used between ICA components and motion signals, and further

work should be done to study other measures.

6.4 Real-time system

The current system relies on a lot of pre-processing and cannot be run online. With modifications to both the motion detection and ICA reconstruction portions of the algorithm, a system could be built to process each frame in under a second.

Instead of computing the Haar-Cascade[**?**] twice for each frame, the motion detection method would be modified to initially compute SIFT[**?**] features and then track optical flow using the Lucas-Kanade algorithm[**?**].

This has applications in neuro-feedback applications and gaming[**?**].

7. CONCLUSIONS

The method presented here shows preliminary results at being able to remove ICA components correlated with facial motion detected through a web-cam. While significant improvements must be made before the method is robust (see Section 6), there are a few benefits offered by the proposed method over similar ones. Namely, this method uses external information (video) that should be available in a non-medical environment using the subject's camera. It also requires no classification training by an expert, which allows subjects to perform EEG tests at home. This could be used for remote diagnosis kits, neuro-feedback, or enhanced learning. Finally, this method may alleviate some concerns over data quality in consumer headsets by allowing software post-processing to make up for the lower quality electrodes.