A Method for Opinion Classification in Video Combining Facial Expressions and Gestures

Airton Gaio Junior
Federal University of Amazonas - UFAM
Institute of Computing - IComp
Email: airton.junior@icomp.ufam.edu.br

Eulanda Miranda dos Santos
Federal University of Amazonas - UFAM
Institute of Computing - IComp
Email: emsantos@icomp.ufam.edu.br

Abstract—**Most of the researches dealing with video-based opinion recognition problems employ the combination of data from three different sources: video, audio and text. As a consequence, they are solutions based on complex and language-dependent models. Besides such complexity, it may be observed that these current solutions attain low performance in practical applications. Focusing on overcoming these drawbacks, this work presents a method for opinion classification that uses only video as data source, more precisely, facial expression and body gesture information are extracted from online videos and combined to lead to higher classification rates. The proposed method uses feature encoding strategies to improve data representation and to facilitate the classification task in order to predict user's opinion with high accuracy and independently of the language used in videos. Experiments were carried out using three public databases and three baselines to test the proposed method. The results of these experiments show that, even performing only visual analysis of the videos, the proposed method achieves 16% higher accuracy and precision rates, when compared to baselines that analyze visual, audio and textual data video. Moreover, it is showed that the proposed method may identify emotions in videos whose language is other than the language used for training.**

I. Introduction

Currently, a large number of people share their opinions, stories and comments through video postings on sites such as Youtube, Vine and Vimeo. Among these, YouTube is certainly the most popular, which receives via upload, more than 300 hours of video per minute. This phenomenon has attracted attention of a large number of companies, investors and consumers, whose focus is developing better opinion mining applications based on online video [1]. For instance, it is possible to obtain sentiment intensity perception involved in people's opinion in order to provide more accurate recommendation and to create more robust profiles of these people.

Most of the literature dealing with opinion analysis use the combination of multimodal data such as text, video and audio. The objective of using more than one modality is to increase the accuracy and precision of the prediction. Many of these studies have proved that data fusion achieves better results when compared to individual modalities [2], [3], [4] and [5]. Nevertheless, the results obtained even employing fusion, in general, are not higher than 80% in terms of precision rate.

Moreover, using information from different data sources can considerably increase the complexity of the model and prevent replication due to several issues, such as: 1) the task of obtaining multimodal data may not be simple, as in the case of text information, which are often transcribed manually; and 2) data from different sources may not be correctly synchronized, which may lead to a noisy fusion process. Challenges particularly important in problems involving video, are: facial expression and body gestures diversity among people from different places and languages; noisy environments, due to videos recorded with different devices; and uncontrolled environments, owing to different backgrounds, illumination and scales. Another drawback to models reported in the literature is the use of commercial software as part of the solution, especially for the automatic feature extraction process.

Given this context, our objective in this work is to take advantage of fusing different modalities, while at the same time overcoming the drawbacks mentioned above. In order to accomplish this objective, we propose a method for opinion classification based on two sources of information extracted from video only, precisely facial expression and body gesture. The choice of these modalities to represent the emotion expressed by a person is inspired by [6], whose authors report that body gesture data can be useful when it is not possible to identify the emotional state of the face, which is normally used as single source of information. Studies presented in [7] and [6] also support this statement. These works cope with emotion recognition by employing both facial expressions and body gestures modalities. Moreover, since the proposed method does not use text nor audio modalities, it is expected to be a language-independent/invariant method.

The method proposed in this article is divided into three phases: **1)** Feature Extraction; **2)** Feature Encoding; and **3)** Decision fusion. In the first step, classic feature descriptors widely popular in the literature are used, such as Motion History Image - MHI, which, in short, represents the movement of an opinion captured in a video sequence; and Histogram of Oriented Gradients - HOG, which extracts information related to the orientation of the gradient of an image. The novelty of our work is in the second phase, by using encoding strategies to improve the representation of features generated by the classic descriptors, and in the third phase, with a fusion approach based on classifier's level of confidence. In our method, the generated feature vectors are coded by techniques based on the Bag of Visual Words Model - BoVW, which is used in computer vision applications and has been adopted as an important image representation model, especially in problems involving human activity recognition, as in [8].

Two encoding algorithms are investigated in this work: Fisher Vector - FV and Vector of Locally Aggregated Descriptors - VLAD, both based on the BoVW concept. In addition, since two sources of information are used, two feature vectors are generated: one with facial expression data and the other with gesture data. Each feature vector is used to represent the data and, thus, to train a classifier. Given that two classifiers are obtained, the last phase of our method involves merging the decision of the two classifiers. The fusion is based on classification confidence, i.e. the classifier with the highest level of confidence is chosen to assign a class to each sample.

The method proposed in this paper was tested using three databases. In addition, as a way of demonstrating portability and language independence, of the proposed model was trained with samples contained in a database composed of opinions expressed exclusively in English, and tested in a database whose opinions are exclusively expressed in Spanish.

This paper is organized as follows: in Section II, we present related work focused on current methods that are used as baselines in our experiments. Then, the proposed method is discussed in Section III. Sections IV and V present experimental protocol and the results obtained respectively. Finally, conclusions and future directions are discussed in Section VI.

II. Related work

One important contribution provided in [5], besides a method for multimodal opinion analysis, is the Multimodal Opinion-level Sentiment Intensity (MOSI) database. The samples contained in this database represent 7 types of opinion intensity: strongly positive, positive, weakly postive, neutral, weakly negative, negative and strongly negative. In this study, the authors seek to understand the pattern of interaction between words (text) and visual gestures. Three modalities are used: text, video and audio. For text data, the simple Bag-of-Words strategy is employed to extract features. In addition, they use more than 32 audio characteristics, including intonation, Mel-Frequency Cepstral Coefficients - MFCCs and Normalized Amplitude Quotient - NAQ, which were obtained by using a collaborative repository of speech analysis for speech technologies [9]. Finally, smile, frown, nod and shake of the head were used to represent video data [10]. Support Vector Machine (SVM) was superior than Deep Neural Network on classification. The accuracy rates attained for text, video and audio were 0.65, 0.61 and 0.57 respectively, when using these sources individually. However, the highest accuracy rate (0.71) was attained using the feature level multimodal fusion, carried out by means of a vector concatenation technique.

Like the previous research, the study presented in [3] addresses opinion recognition on combining text, audio and video. An important contribution of this research is the Multimodal Opinion Utterances Dataset - MOUD. This database provides samples divided into three different classes of opinion: positive, negative and neutral. Audio features, such as pause duration, pitch variation, volume media and voice strength were automatically extracted using the OpenEAR software. Information related to video were captured using the commercial software called Okao Vision[1], which automatically returns smile intensity and the direction of the look according to horizontal and vertical angles. For text information, the authors used a Bag-of-Words approach with the transcriptions of sentences to construct a vocabulary in order to extract features. The multimodal fusion was feature-based and conducted by simple vector concatenation. The classification algorithm used was SVM with linear kernel. Again, given that fusing the three modalities achieved 0.75 as accuracy rate, this result outperformed single modalities rates: 0.649 for text, 0.610 for video, and 0.467 for audio. In addition, the authors investigated the portability of the method using a second database composed of 37 opinions about cell phones captured from Expotv.com. The multimodal fusion reached 0.648 as accuracy rate, which is higher than rates for text (0.540), video (0.540) and audio (0.486).

In [4], a dataset proposed in [2] for the analysis of multimodal opinion is investigated. The database, called Youtube Dataset, is composed of Youtube videos divided into three classes of opinions: positive, negative and neutral. Using the Luxand FSDK 1.72[2] commercial facial recognition software, the authors extracted 66 face points throughout video frames and calculated the mean distance of these points to form the video modality final feature vector. Likewise, audio features, such as tone and voice strength, were also extracted from audio segments using the OpenEAR software. Text information were obtained from transcriptions according to a heuristic for sentiment analysis. Among the machine learning algorithms tested by the authors there are: Naive Bayes, SVM, Extreme Learning Machine (ELM) and Artificial Neural Network. ELM was better than the other methods. The results for individual classifications attained for text, audio and video in terms of accuracy were 0.619, 0.652 and 0.681 respectively. The authors performed the feature-level multimodal fusion, reaching the highest accuracy reported at 0.772, while for the decision-level multimodal fusion, 0.753 of accuracy rate was attained.

In a different way, focusing on reducing the opinion classification process complexity, our work deals with only two modalities: face expression and body gesture, both extracted from video as a single data source. Another advantage on using face expression and body gesture is to allow an opinion classification approach less sensitive to the language spoken by the person who expresses the opinion. In addition, we have adopted the use of encoding methods which, according to the literature, have not been applied in the context of multimodal data fusion with focus on opinion classification. However, this kind of technique is widely used in problems involving human activity recognition [8]. Finally, data fusion is conducted using a decision-level fusion strategy based on the classification confidence, as described in the next section.

[1] Available at the website: https://plus-sensing.omron.com

[2] Available at the website: https://www.luxand.com/index.php

III. Proposed Method

Figure 1 shows the architecture of the proposed method. As can be seen in this figure, the process illustrated in this architecture is traditional in standard pattern recognition solutions. Initially, since the method employs information from faces and body gestures of each individual, the input videos are provided to the modules of face detection, and feature extraction. In this work, the face extraction module is performed using HOG. Likewise, the input videos are provided to the body gesture feature extraction module, which is conducted by combining two descriptors: MHI and HOG. Thus, the dimensionality of feature vectors obtained from faces and body gestures is reduced by PCA (Principal Component Analysis). Then, a feature encoding strategy is employed to generate feature representations more likely to be better for classification.

Two encoding algorithms are investigated in this paper: VLAD and FV. Then, data described using the representations generated by the encoding module are used to train a machine learning algorithm. Among the many possibilities available in the literature, we use SVM without parameter settings, i.e. in its linear kernel version, and providing a posterior probability as output. It is important to mention that encoding algorithms are expected to perform well even with simple linear classifiers, since they provide an embedding of the local descriptors in a higher-dimensional space which is easier to deal with by linear classifiers [11]. This is an advantage, since linear classifiers are efficient for both training and evaluation. Finally, the results of classifying face and body modalities are combined with a fusion technique. The details about the proposed method are discussed in the next section.

A. Input Data

We believe that our method can be robust to classify opinions expressed in real-world videos with varying qualities and sizes, acquired with different types of equipment and with differentiated backgrounds and illuminations. However, it is important to highlight that the video databases investigated in this work present as main characteristic the existence of only one person expressing an opinion on any subject. The person is invariably positioned facing the camera. In this way, it is possible to see the whole person's face.

B. Face modality

The classic literature shows, in several studies, that the following main steps need to be carried out in order to properly use the face modality: face detection and feature extraction [6].

1) Face Detection: Due to its efficiency and velocity when applied for face detection, the Viola-Jones algorithm was chosen for this task [12]. Initially, video frames are converted from RGB to grayscale, then they are used as input parameter for the algorithm that tracks the face and returns the face bounding box. This kind of bounding box allows the segmentation of the area of interest. Again we employ Viola Jones, this time with specific parameters, to return the limits of the mouth, nose and eyes, obtained from the image face previously tracked as input to the algorithm. It is important to mention that Viola-Jones is not robust for unconstrained environments, which must be dealt with using other approaches [13].

Given that the videos are not controlled, there may be varied angles of the faces of people who express their opinions spontaneously. In such cases, it is possible that the detection algorithm cuts blocks of slightly different sizes. Focusing on overcoming this problem, blocks are resized after detection, as follows: eyes (25×95 pixels); nose (35×42 pixels); and mouth (34×55 pixels). Then, blocks are normalized according to the mean and standard deviation in order to mitigate noise. Finally, a normalized image of these blocks is obtained and used as input to the feature extraction module.

2) Feature Extraction: In this module, blocks generated in the previous phase are submitted as input to the HOG descriptor [14], which is based on the evaluation of standard local histograms of the orientation image gradient in a dense grid. Taking into account the small size of the blocks (eyes, mouth and nose), we define a window with a 4×4 pixels size as a parameter for the descriptor. On the one hand, smaller sized windows provide more information, on the other, they generate vectors of larger sizes. Finally, as a result of this step, a set of features extracted from the blocks are concatenated in a single and long vector to be used in the next step of the method. In this paper, eyes, mouth and nose features were concatenated in a single vector, generating 9,253 features.

C. Body gesture modality

The input videos are also provided to the body gesture feature extraction module. However, in this case, there is no segmentation of specific area in the frame. We simply consider the entire frame as input information for the process. As the videos have different sizes, we resized all frames to 360×480 pixels. In this work, body feature extraction was performed using 2 descriptors: MHI and HOG. While MHI represents a motion sequence of the body in the video, HOG extracts the orientation of the gradient distribution from the MHI output [14]. These steps are detailed in the next subsections.

1) MHI: In [15], authors report that the intensity of the pixel in an MHI image represents the history of movement at that point, i.e. brighter values correspond to a larger and more recent movement. In this work, MHI is used to represent the motion body of a video stream. The process of generating the MHI image is quite simple. The body video sequence is converted from RGB to grayscale. Then all n frames are captured by calculating the difference image of all frames, resulting in a set $S = \{1, 2, 3, ..., n-1\}$ of difference images. Here, 5 was experimentally defined for the intensity threshold, which controls the intensity of the movement records. Finally, a results weighted sum of the filtered images is performed, considering 255 as the intensity scale, thus obtaining a single image. Figure 2 shows an example of the generation of an image motion representation, which was captured from a video stream. This process provides a gradient image synthesizing in space and time the movement of an expressed opinion. The output image is then used as data input for the next descriptor.

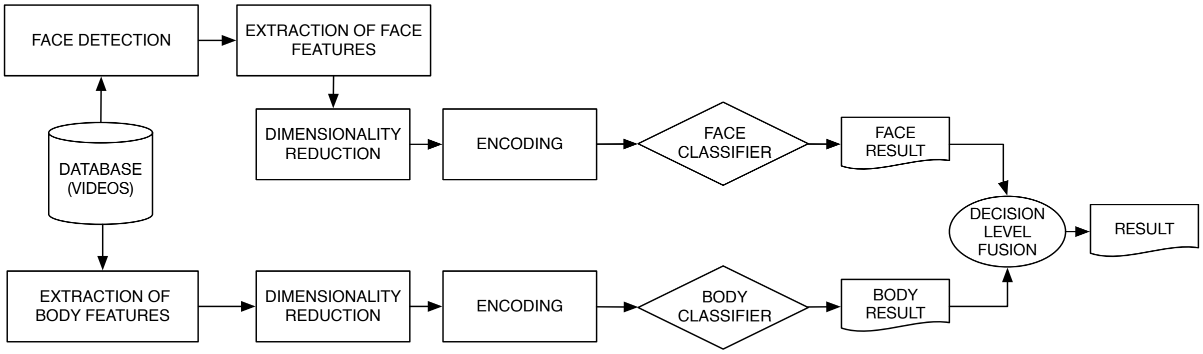


Fig. 1. The general architecture used for multimodal opinion recognition. PCA is employed for dimension reduction; and encoding algorithms are employed using data provided by face and body descriptors. A classification method is then trained to learn to classify opinions using face and body modalities individually. Finally, a fusion strategy is employed to combine the output of the two individual classifiers to produce a single classification assignment for the input data.

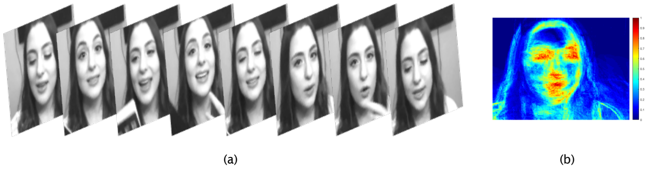


Fig. 2. Using MHI to extract body features. (a) Sequence of frames showing a positive opinion about a book; and (b) MHI representation with an applied jet palette. The lighter gradient values correspond to a greater body movement, darker values represent less body movement.

2) HOG: In order to complement the process of extracting body features with more information, this work employs HOG descriptor in the representation of the movement generated in the previous step. Figure 3 demonstrates this feature extraction process performed using the MHI generated image. In the same way as in the previous modality, the window size was set to 4×4 pixels when using HOG. Due to the small size of the images, we chose the lowest value for HOG algorithm, in order to extract more features of the face and gesture of the body. In this paper, 381,277 of body features were extract.

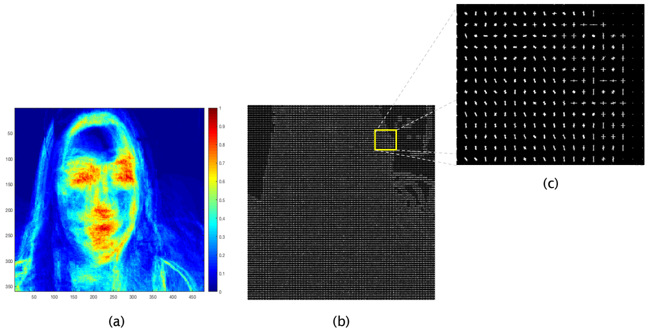


Fig. 3. HOG feature extraction process using the MHI generated image as input. (a) MHI representation with an applied jet palette; (b) Visual representation of HOG extracted from the MHI output; and (c) Details showing the orientations of the gradient generated by HOG.

D. Dimensionality Reduction

Due to the fact that the feature vectors extracted in the previous modules are high dimensional, PCA was applied in this paper to reduce both feature vectors dimensionality. We have defined a 95% threshold of principal components as a guarantee of maintaining the representativeness of the information extracted. After dimensionality reduction we obtained 1,098 features of body gesture and 697 features for face. As a result of this module, lower dimensional and high representative feature vectors are generated, in order to be used in the next module of the proposed method.

E. Encoding

FV and VLAD encoding methods are both based on the BoVW concept. In general, a traditional BoVW framework contains local feature extraction, visual dictionary generation based on a clustering algorithm, such as K-means and Gaussian Mixture Models - GMM, besides the feature encoding. Briefly, the coders use the local features to generate a general signature, causing a better spatial separation between the classes, simplifying the work of the classifiers. These two algorithms are described in the next sections.

1) Fisher Vector - FV: According to [16], FV is a representation of the image obtained through local image features. It is often used as a global image descriptor in visual classification. Before encoding, it is necessary to create a visual vocabulary combining local information extracted from the training videos. A particular characteristic of the FV encoding method is that the clustering algorithm used to generate the dictionary of visual words should be exclusively GMM. The objective of GMM is to identify the presence of subpopulations contained in a dataset. This clustering algorithm need a single parameter to be set: k - number of clusters. The parameter k is data-dependent, so we performed some experiments to set its value, as described in section IV-B.

Let $I = (x_1, ...x_N)$ be a set of D dimensional feature vectors extracted from an image, for instance using HOG. Let $\theta = (\mu_k, \Sigma_k, \pi_k : k = 1, ..., K)$ the GMM parameters adjusted and associated to each vector x_i for a value of k in the mixture as a weight given by the posterior probability:

$$q_{ik} = \frac{exp\left[-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right]}{\Sigma_{t=1}^{k} exp\left[-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right]} \qquad (1)$$

For each value k, consider the mean vectors and covariance deviation.

$$u_{jk} = \frac{1}{N\sqrt{\pi_k}}\Sigma_{i=1}^{N} q_{ik}\frac{x_{ji}-\mu_{jk}}{\sigma_{jk}}, \qquad (2)$$

$$v_{jk} = \frac{1}{N\sqrt{2\pi_k}}\Sigma_{i=1}^{N} q_{ik}\left[\left(\frac{x_{ji}-\mu_{jk}}{2}\right)^2 - 1\right] \qquad (3)$$

Where $j = 1, 2, ..., D$ refers to the dimensions of the vector. The FV encoding of the image I is the stacking of vectors u_k and v_k for each of the values K in the Gaussian mixtures.

This technique encodes the difference between the data according to the number of clusters k and applies derivative operations on the probability based on the clusters spacial distribution. One advantage of VF when compared to other encoding methods is that the dictionary generated by GMM captures both first order and second order statistics. By providing to GMM the vectors generated after the dimensionality reduction step and the number of clusters, valuable information such as the clusters centroids and the covariance matrix are obtained. Then, VF uses this information to encode the difference between the clustered descriptors and the vocabulary. This is attained by applying derivative operations on the parameters of the vocabulary distribution probability [17].

The FV encoding method usually produces high performance and relatively few visual words, even if simple linear classifiers are used. Its disadvantage is that the result output vector after encoding will have a fixed size, precisely $2\times k\times d$, where the number of clusters indicated in the previous step is the number of dimensions of the feature vector.

2) VLAD: It is a version of VF that only maintains first-order statistics [8]. According to [17], VLAD accumulates the residual of each local feature in terms of its assigned visual word. So, it combines each local feature with its closest visual word. Finally, for each cluster, it stores the sum of the differences of the descriptors assigned to the cluster and centroid of the cluster. The dictionaries are generated by K-means. The size of the output vector encoded with VLAD is $k \times d$, where k is the number of clusters indicated in the previous step and d is the feature vector dimension.

F. Classification and Fusion

Due to the information gain in terms of spatial distribution that encoding methods produce, it is not necessary to employ complex classifiers to achieve good results, leading to simpler learning process. For example, in our experiments we use SVM with linear kernel without tunning additional parameters. VLAD and FV encoded feature vectors are provided as input to classifier training and testing. Thus, two different classifiers are provided, one trained with face data and another classifier trained with body gestures data, generating isolated results.

According to the literature, multimodal fusion may be undertaken at feature-level or at decision-level. However, as in our method two classifiers are trained, it is necessary that the two generated models be combined at decision-level, so that only one class is assigned to the input data. Among the various forms of classifier fusion available, a fusion strategy based on classification confidence, which is measured as the classifier posterior probability, is used in this work. Taking into account the probability values generated by the classifiers, the sample is assigned always to the class with the highest probability based on the following rule: **1)** - when the two classifiers predict the same class for the input instance, this class is assigned to the instance; **2)** when the two classifiers diverge, the classifier with the highest probability is chosen to assign the class to the input instance. In this way, the chosen classifier is most likely to be the correct one for classifying the input instance.

IV. Experimental Protocol

In this section the three databases used in the experiments are described. Next, FV and VLAD parameter settings necessary for the generation of dictionaries used to encode body gesture and face modalities is presented.

A. Databases

The proposed method was investigated using three databases: Youtube Dataset, developed in [2]; Multimodal Opinion-level Sentiment Intensity - MOSI created in [5]; and Multimodal Opinion Utterances Dataset - MOUD, described [3]. Although not controlled databases, all these datasets are labeled and contain video, audio, and text files with the transcriptions of speech. The latter is obtained manually.

1) Youtube Dataset: It is composed of 47 videos acquired directly from the YouTube site. It deals with a variety of topics, from opinions on products, religion or even on political positioning. In all, there are 20 females and 27 males, aged between 14 and 60 years, with different ethnic origins and who express their opinions in English. The database can be requested from the electronic address *http://projects.ict.usc.edu/youtube/*. Videos are on Moving Picture Experts Group - MPEG-4 format with 360×480 pixels as default size. The videos last from 2-5 minutes, and each video belongs to one of three classes: positive (1), negative (-1) and neutral (0). Despite the small number of samples, authors have attempted to balance the distribution among the classes: 32% positive, 25% negative and 43% neutral.

2) MOSI Dataset: Like the database described above, MOSI is composed of videos that were collected from Youtube focusing on popular video-blogs used by many people to express opinions about different subjects. Labled according to the intensity of the opinion, it presents a total of 7 defined classes: strongly positive (+3), positive (+2), weakly positive (+1), neutral (0), weakly negative (-1), negative (-2) and strongly negative (-3). However, in this work, we use only videos labeled with the 3 basic classes of opinion: positive, negative and neutral. According to Zadeh et al. [5], there are challenges in this collection, since videos are recorded with different configurations, some people used professional quality devices, while others, less professional devices. In addition, videos are recorded at different distances and with different background and lighting conditions. The MPEG-4 format was retained from the original site, thus, their original sizes were preserved, resulting in videos with different sizes. The durations of the videos range from 2-5 minutes.

A total of 89 different people were selected in the videos, 41 women and 48 men with ages between 20 and 30 years of different ethnicities, all expressing their opinions exclusively in English. Although the database has a total of 89 videos, the authors divided the videos into smaller segments according to the intensity of the opinion, resulting in a total of 1,298 videos. As in the database described in subsection IV-A1, MOSI maintains a balanced class distribution: 39% for the positive class, 35% negative and 26% for neutral. The database can be requested at *https://goo.gl/forms/vFfFCdP2Jua8Wwtm2*.

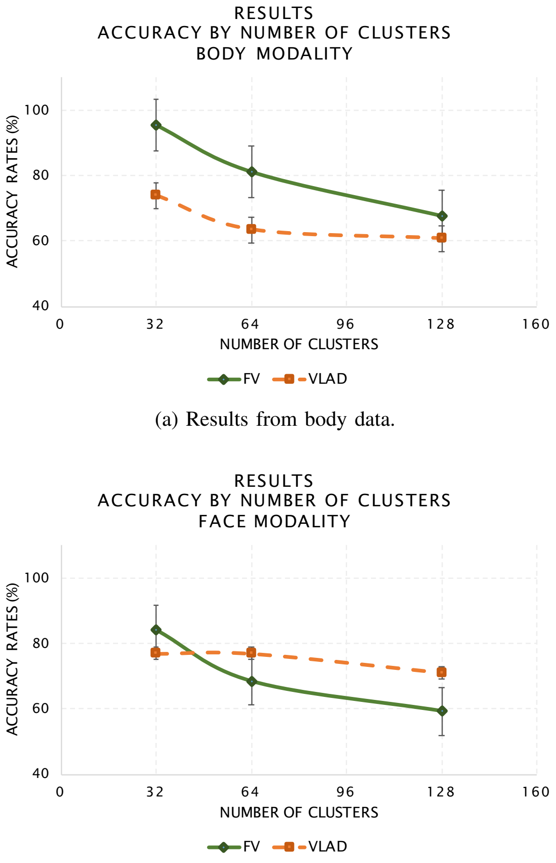
3) MOUD Dataset: It consists of a set of 105 videos collected from Youtube, in which people express their opinions mostly about movies, cosmetic products and books, exclusively in Spanish. The final video dataset includes 21 males and 84 females randomly selected, with ages ranging from 15 to 60 years, from Spanish-speaking countries, such as Mexico, Spain or South America countries. This dataset is publicly available at *http://lit.eecs.umich.edu/*.

All videos were converted to the MPEG-4 format with a standard 352×288 pixels size. Video durations range from 2-8 minutes. It is important to note that there is an imbalanced distribution between classes, with the neutral class having only 10% of the total number of samples. The original 105 videos were segmented into smaller videos according to the opinion statement, resulting in a total of 482 samples.

B. Parameter Settings

According to the encoding methods employed in our research, both FV and VLAD require a pre-processing owing to K-means and GMM clustering algorithms used to generate visual dictionaries. The objective of this step is to group n observations among k clusters, where each observation belongs to the cluster closest to the center. As mentioned before, VLAD employs K-means while GMM is used in the FV encoding method.

As a way of defining the value of k for each encoding method, we perform tests using the MOSI database, since it has the highest number of samples among the datasets investigated. Three values were investigated: 32, 64 and 128. Figure 4 shows the results obtained for body (4a) and face (4b) modalities by using FV and VLAD with different number of clusters. The results show that, for both body or face modalities, applying any encoding algorithm, the accuracy rates attained with 32 clusters were higher than the results achieved when larger number of clusters are used. Thus, the value of 32 for parameter k was taken as the default value in all other experiments in this work.



(a) Results from body data.

(b) Results from face data.

Fig. 4. FV and VLAD encoding methods accuracy rates obtained for body and face modalities by varying the number of clusters.

V. Experiments and Results

In this section we describe the results achieved by the proposed method in the databases investigated. These results are compared to the results obtained by baselines. Finally, we present an analysis of the performance of the proposed method when trained with videos of one language and tested with videos of another language. The objective of this second series of experiments is to evaluate whether or not our method is invariant to language. To do so, the MOSI dataset is used to train the learning model and MOUD is used to test it. SVM classifier with linear kernel was employed in all experiments.

A. Youtube Dataset Results

For purposes of comparing the results obtained in this database, we used the 10-fold cross-validation strategy, based on the baseline experimental protocol conducted in [4]. Accuracy is used as the performance metric.

As already mentioned in Section II, the research developed in [4] performs multimodal fusion of video, audio and text, in addition to dealing with two different fusion approaches:

feature-based and decision-based fusions. The results presented in Table I demonstrate that, for both baseline and our method, multimodal fusion achieved a better result when compared to the results attained with individual modalities. Taking into account that our method deals only with video data, the best result achieved when comparing face and body gesture-based classifiers is reported in all tables in this section, as video individual modality accuracy rate.

TABLE I
RESULTS OBTAINED BY THE PROPOSED METHOD ON YOUTUBE DATASET COMPARED TO THE BASELINE RESULTS.

Method	Video	Audio	Text	Fusion
Baseline	0.68	0.65	0.61	0.78
VLAD	0.57	-	-	0.60
FV	0.77	-	-	**0.84**

The highest accuracy achieved by the baseline was 0.78, obtained with feature-based fusion. However, it can be seen in this table that this rate is lower than the rate we have reached with our method (0.84), which only employs face and body gesture data. This accuracy rate was attained when FV was applied as encoding method. If we compare the accuracy obtained by the baseline with video-only data (0.68) to face (0.77) and body gesture (0.71) individual modalities reached by our method with the FV encoding, our method outperforms these results. On the other hand, VLAD encoding strategy presented worse results compared to the baseline.

B. MOSI Dataset Results

The experiments conducted with MOSI were performed by 5-fold cross-validation, similar to the baseline developed in [5]. As in the previous experiments, the authors of the baseline investigated opinion classification using video, audio, and text. In addition, they employed a decision-based fusion technique and compared two types of classifiers: SVM with linear kernel and Deep Neural Network, however, the best reported results were achieved using SVM.

Table II compares the baseline results to our method's. It is possible to note that, again the multimodal fusion achieved higher accuracy compared to the results obtained when using the isolated modalities, for both baseline and the proposed method. Moreover, in this series of experiments, SVM trained with data representation provided via FV encoding method, also outperformed VLAD, since it reached 0.94 as accuracy rate against 0.83 from VLAD. When compared to the baseline, all two versions of the proposed method attained higher accuracy rates, since the baseline reached 0.61 for video-only data and 0.71 with fusion.

TABLE II
COMPARING THE PROPOSED METHOD TO THE BASELINE ON MOSI DATASET.

Method	Video	Audio	Text	Fusion
Baseline	0.61	0.57	0.65	0.71
VLAD	0.77	-	-	0.83
FV	0.92	-	-	**0.94**

C. MOUD Dataset Results

The results of this group of experiments are compared to the work conducted in [3], where linear SVM was employed for classifying video, audio and text by 10-fold cross-validation. In this database, the results are different from previous ones, since VLAD was better than FV. It is important to emphasize that FV attains better performance when dealing with the body gesture modality. Therefore, our hypothesis for VLAD superiority here is that the FV encoding process for body gesture data was less effective, since we observed a large number of video frames with a short execution time (less than ten seconds). This behavior may have contributed to decrease the representation of the MHI descriptor, and consequently of HOG's.

We show in Table III, the comparison of the results obtained by the baseline and by our proposed method. Both methods again presented results with higher accuracy rates when fusing modalities, instead of using individual data sources.

TABLE III
COMPARING THE PROPOSED METHOD TO THE BASELINE ON MOUD DATASET.

Method	Video	Audio	Text	Fusion
Baseline	0.61	0.47	0.65	0.75
VLAD	0.93	-	-	**0.95**
FV	0.76	-	-	0.80

In terms of the three modalities investigated separately by the baseline, the modality that obtained the highest accuracy was that of text with 0.65, even though it was below the face and body modalities used in our method, regardless of the type of encoding algorithm employed. The best result reached by the baseline was obtained with decision-based fusion, 0.75, also worse than the two versions of method: 0.80 with FV and 0.95 with VLAD encoding.

D. Training and testing in different languages: MOSI versus MOUD

As a way of evaluating in more depth the method we propose in this work, and also, demonstrating portability and language independence, we carried out experiments using samples of MOSI (English) database for training a model and samples of the MOUD (Spanish) database for testing. In all, there were 1,298 training and 482 test samples. The SVM classifier with linear kernel was used in the experiments and accuracy was again the evaluation metric employed.

Figure 5 shows that the results obtained using VLAD were very low accuracy rates, being below 0.50. On the other hand, FV enabled better results to be obtained, both with body gesture (0.74) and face (0.75) data, but again the best result was achieved with the fusion of face and body modalities with 0.82 accuracy. Therefore, these results indicate that the proposed method allows portability regardless of the language spoken in the training videos, because it is possible to use the method to classify opinion in video whose language is different from the language used for training.

In all experiments the results of face and body modalities fusion overcame the individual modalities. This was also

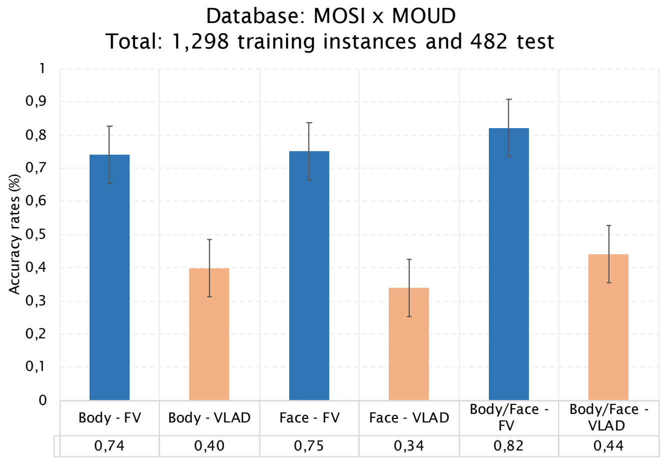


Fig. 5. Test results on evaluating language independence.

observed in this section. In general, FV encoding achieves the best results with body gesture data, while VLAD, is more prone to better work with face data. This behavior contributed to the diversity of the results and, consequently, there was an improvement due to the encoding methods. Our results confirm the study [8], which demonstrates the superiority of FV over the other encoding algorithms.

It can be concluded, therefore, that the proposed method presents some advantages over current solutions, since even using only body gesture and face information, exceeds in average 16% baselines based of fusing video, text, and audio. In addition, the proposed method employs classical and publicly available feature extraction techniques, while the baselines use proprietary software. In general, data collected from body gesture and face produced enough diversity to lead to a better decision-based fusion performance. Finally, it is possible to use the proposed method to identify emotions in videos with languages other than the language used in the training dataset.

VI. Conclusion

The aim of this work was to develop a method of multimodal opinion classification based on combining information extracted from facial expressions and body gestures from online video.

The method proposed in this work was tested in three databases and compared to three different baselines, surpassing them in approximately 16% of accuracy. During the analysis of the results, it was possible to notice that the use of encoding methods significantly improves the accuracy rates of the proposed method, even when a less robust classifier as SVM with linear kernel is used, and without parameter settings. In general, FV achieved better results when compared to VLAD. Face and body modalities produced complementary information when providing the classifiers' level of certainty. This fact significantly helped the multimodal fusion to achieve the best results when compared to the results attained when using isolated modalities, since it provided diversity for the selection process, which is an advantage for fusion rules.

Due to the fact that we do not use audio and text data, we consider that the proposed method allows portability regardless of the language spoken in the video. In order to reinforce this statement, experiments have shown that it is possible to use our method to classify opinion in video whose language is different from the language used for training it. In this context, the proposed method still obtained 82% as accuracy rate.

As future research we intend to deepen the analysis incorporating other degrees of emotion expressed in opinion videos as: weakly positive, weakly negative, strongly positive and strongly negative.

References

[1] A. Zadeh, "Micro-opinion sentiment intensity analysis and summarization in online videos," pp. 587–591, 2015.

[2] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," pp. 169–176, 2011.

[3] V. P. Rosas, R. Mihalcea, and L. P. Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 38–45, May 2013.

[4] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, Part A, pp. 50 – 59, 2016.

[5] A. Zadeh, R. Zellers, E. Pincus, and L. Morency, "MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *CoRR*, vol. abs/1606.06259, 2016.

[6] H. Gunes, C. Shan, S. Chen, and Y. Tian, "Bodily expression for automatic affect recognition," *Advances in Emotion Recognition*, no. July, pp. 1–34, 2013.

[7] A. Panning, I. Siegert, A. Al-Hamadi, A. Wendemuth, D. Rosner, J. Frommer, G. Krell, and B. Michaelis, "Multimodal affect recognition in spontaneous HCI environment," *2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012)*, pp. 430–435, 2012.

[8] X. Peng, L. Wang, and X. W. andmethods for action recognition: Comprehensive study and good practice," *CoRR*, vol. abs/1405.4506, 2014.

[9] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep x2014; a collaborative voice analysis repository for speech technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 960–964.

[10] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of eyes for eye-shape registration and gaze estimation," *CoRR*, vol. abs/1505.05916, 2015.

[11] P. F. M. T. Sanchez, J. and J. Verbeek, "Image classificationwith the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, pp. 222–245, 2013.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[13] F. Zavan, N. Gasparin, J. Batista, L. Silva, V. Albiero, O. Bellon, and L. Silva, "Face analysis in the wild," in *Tutorials of the 30th Conference on Graphics, Patterns and Images (SIBGRAPI'17)*, 2017. [Online]. Available: http://sibgrapi2017.ic.uff.br/

[14] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pp. 886–893, 2005.

[15] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, "Hierarchical filtered motion for action recognition in crowded videos," *Systems, Man, and . . .*, vol. 42, no. 3, pp. 313–323, 2012.

[16] F. Perronnin and C. Dance, "Fisher Kernels on Visual Vocabularies for Image Categorizaton," *Proc.* {*CVPR*}, 2006.

[17] P. T. Biliński, "Human action recognition in videos," Theses, Université Nice Sophia Antipolis, Dec. 2014.