ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 2, No.14 , March 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Affective Video Retrieval Based on Dempster–Shafer Theory

Shahla Nemati[1] and Ahmad Reza Naghsh–Nilchi[2]

[1] Department of Computer Architecture, University of Isfahan, Faculty of Computer Engineering
Isfahan, Iran
nemati@eng.ui.ac.ir

[2] Department of Artificial Intelligent, University of Isfahan, Faculty of Computer Engineering
Isfahan, Iran
nilchi@eng.ui.ac.ir

Abstract

Affective video retrieval systems are designed to efficiently find videos matching the desires and needs of Web users. These systems usually use fusion strategies to combine information from different modalities aiming at understanding others' affective states. However, common fusion strategies used for affective video retrieval, neither were designed for this task, nor have any theoretical foundation. In order to address this problem, a novel fusion method based on the *Dempster–Shafer theory of evidence* is suggested. This method is utilized to combine audio and visual information contained in video clips. In order to show the effectiveness of the proposed method, experiments are performed on the video clips of DEAP dataset using two popular machine learning algorithms, namely SVM and Naïve Bayes. Results reveal the superiority of the proposed approach in comparison with the existing fusion strategies using both algorithms.

Keywords: Affective video retrieval, multimodal, fusion, Dempster–Shafer theory of evidence.

1. Introduction

Multimedia systems provide on−demand access to enormous volumes of high quality content [1]. In order to effectively achieve this goal, multimedia content analysis (MCA) techniques are employed. MCA aims at developing models for bridging the semantic gap between low−level features and the semantics carried by multimedia contents. There are essentially two approaches to MCA: the cognitive approach and the affective approach. In the cognitive approach, a given multimedia content is analyzed in terms of the semantic features of a scene such as location, characters and events. The main goal of the affective approach, on the other hand, is to predict viewers' emotional reactions in response to an input multimedia content. Although most of the MCA research efforts were focused on cognitive methods, but the importance of the affective approach has been rapidly increasing due to the growing awareness of its role in personalized multimedia recommendation [2].

There are several everyday multimedia applications for affective analysis such as augmenting video delivery websites with more convincing recommendations and enabling parents to better manage what their children watch by knowing the emotional contents of videos [3].

However, due to the enormous volume of multimedia contents on the web, finding appropriate video contents matching the desires and needs of users remained a challenging problem. In order to overcome this affective analysis problem, most researchers have followed single modality approach [4, 5]. Nevertheless, recent affect detection studies show that using just one modality or channel is not sufficient to accurately and consistently detect human affective states. Therefore, multimodal affect recognition approaches are becoming increasingly popular [5]. These approaches typically use a combination of different modalities used by humans to understand others' affective states.

Multimodal approaches allow for more reliable estimation of the human emotions by considering more sources of information. Moreover, they increase the confidence level of the results and decrease the ambiguity with respect to the estimated emotions from separate channels. Nevertheless, the complexity of multimodal affect detection is higher than the complexity of the unimodal approaches. This is due to the fact that there are usually some ambiguity and correlation among different informational channels [6]. In order to overcome this complexity, computer reasoning techniques and machine learning methods are usually applied to a combination of modalities or channels. Typical examples of multimodal aggregation includes audio−visual, speech−text, dialog−posture, face−body−speech, speech−physiology, face−physiology, and multi−channel physiology [5].

Existing multimodal video retrieval methods either employ low–level audio–visual features or construct high–level

attributes from low–level ones. Preserving global relations in data is the main advantage of using high–level features, but it has been shown that creating such features is time consuming and problem–dependent [7]. Therefore, it is preferred to use a combination of low–level features [1]. In the current study we also employ low–level audio–visual features in the proposed multimodal affective video retrieval system.

Regardless of which type of features is used, the fusion algorithm for aggregating affective information from different modalities is the main part of multimodal systems. However, to our knowledge, existing fusion methods for affective video retrieval neither were designed for this task, nor have any theoretical foundation. In order to address this problem, a novel fusion method based on the *Dempster–Shafer (DS) theory of evidence* is suggested in this paper [8]. This theory is applicable to the problem of multimodal affective video retrieval, since the decision made based on audio and visual modalities is an evidence for the affect category of the video.

In order to assess the effectiveness of the proposed method, experiments are carried out on the video clips of DEAP dataset, a multimodal dataset for analyzing human affective states [9]. The main novel contributions of this paper are as follows:

- We propose a fusion method based on the Dempster−Shafer theory of evidence for affective video retrieval.
- We investigate the effect of applying different fusion methods for affective video retrieval.
- We adopt the combination rule of the Dempster−Shafer theory of evidence for the fusion of audio−visual contents.
- We compare the performance of employing fusion methods in both feature−level and decision−level.

The remainder of the paper is organized as follows. Section 2 reviews background and related works; Section 3 illustrates the methodology and the proposed system; Section 4 reports experimental results and finally Section 5 sets out conclusion and future work.

2. Literature Review

2.1 Affect Representation

In the literature, the terms *affect* and *emotion* have been used to show the same concept. However, affect is usually used to describe both long–term (i.e., *Personality* and *Mood*) and short–term (i.e. *emotion*) aspects of human feelings [10]. In the current study, the emotional aspect of affect is considered because according to the movie presented to the audience they express short–term reactions.

Discrete emotion psychologists argue that there are six or more basic affects [11]. Some emotion researchers, on the other hand, believe that there is a correlation between affective states and hence, emotion is better expressed in a dimensional manner, rather than in terms of some discrete emotion categories. For instance, a three−dimensional valence (i.e. positive versus negative affect), arousal (i.e. low versus high level of activation), and dominance (the degree of control over the emotion) space (called V−A−D space) developed by Russell and Mehrabian, from which valence and arousal are among the most accepted dimensions [10]. A simplified model based on V−A−−D space is the two−dimensional V−A space in which the underlying dimensions are valence and arousal. Fig. 1 shows three−dimensional V−A−D and two−dimensional V−A spaces, respectively.

As depicted in Figure 1, only some parts of these spaces are relevant. The two−dimensional model can represent different emotional states [11]. Moreover, it has been shown that basic common emotions (e.g., fear, anger, sadness, etc.) can be represented as different areas on V−A or V−A−D coordinates [2]. In this study, we also follow the dimensional approach considering the valance−arousal dimensions.

2.2 Visual and Audio Analysis

Visual analysis is the main part of affective video retrieval systems. For this task, different visual features may be extracted from video content. For instance, *motion* and *color* features were used to represent arousal and valence [12] while, *lighting key* and *color energy* were used for emotional video tagging [13].

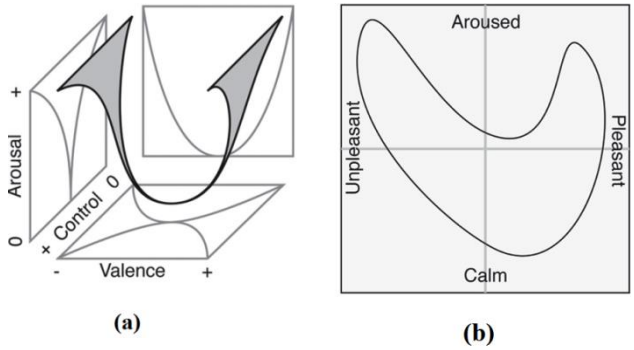

Fig 1. Relevant areas of (a) three−dimensional emotion space and (b) two−dimensional emotion space [11].

Other low–level visual features such as *color activity*, *color weight*, *color heat*, RGB histogram and lightening key were also exploited [3]. However, it has been shown that the most common low–level visual features for multimedia content analysis are motion, color and lighting key [7, 11].

The next important modality for multimedia content analysis is speech that conveys both linguistic (explicit) and paralinguistic (implicit) affective information. Similar to visual analysis, different audio features are also used for affect recognition. For example, Yazdani et al. used *zero crossing rates* (ZCR), *Mel frequency cepstrum coefficients* (MFCC), and *delta MFCC* to specify emotion in music video clips [11]. More recently, Acar et al. used 13−dimensional MFCC as low−level audio features [7]. However, the most common audio features are *pitch*, ZCR, MFCC, and energy.

2.3 Multimodal Fusion

As discussed earlier, fusion algorithm plays an important role in multimodal approaches. Fusion methods are used to integrate affective information from different sources, probably on different time scales and measurement values. Fusion strategies are classified into feature−level (early integration) and decision−level (late integration) categories [6]. In the feature−level, feature vectors are first extracted from the respective modalities and then they are combined together before the classification stage. Finally, a classifier is used to learn the properties of the joint observation. This approach has the advantage of taking into account the correlation between different modalities. However, it does not generalize well for modalities with different temporal characteristics (e.g. speech and gesture inputs) and hence, is more applicable for closely coupled and synchronized modalities (e.g. speech and lip movements). Moreover, in order to train the classifier, large amounts of labelled data must be collected due to the high dimensionality of features vectors [14]. Decision−level fusion methods, on the other hand, first classify the feature vectors of each modality separately and the then, combine the classifiers' outputs into a final decision. Designing optimal strategies for decision−level fusion has remained an open problem in the literature [14]. Existing works on multimodal emotion recognition have considered both feature−level and decision−level fusion. Different combinations including face−body, face−speech, face−physiological signal, face−voice−body, speech−physiological signal, and speech−text have been tested [13]. For example, Scherer and Ellgring combined facial, vocal features, and body movements (posture and gesture) to discriminate among 14 emotions [15]. More recently, Castellano et al. try to detect emotions by monitoring facial features, speech contours, and gestures [16]. However, there are very few systems that have investigated multimodal affect detection. These primarily include studies that combined physiological sensors and those which combined acoustic−prosodic, lexical, and discourse features [4, 17] .

3. Methodology

Fig. 2 shows an overview of the proposed system. The input to the proposed system is a typical music video clip whose affect category should be determined. As discussed earlier, in the current study we follow the dimensional approach for specifying the affective states of viewers after watching the input music video clip. To this aim, the well−known V−A space is considered for showing the output of the system. Specifically, the output of the proposed system is a label showing one of the four quadrants of the V−A space.

In the classification modules a supervised strategy is used. The affect labels provided in the DEAP dataset are used for training audio and visual classifiers. In the DEAP dataset, the arousal and valence values of each video clip are specified with an integer in the range of 1 to 9. However, the purpose of the current study is to classify video clips into one of four quadrants of the V–A space. Therefore, arousal and valence values are mapped to four quadrants of the V–A space as follows. Each quadrant is specified with one of the following labels: *negative–high* (NH), *negative–low* (NL), *positive–high* (PH), and *positive–low* (PL), respectively. Then, for arousal, if the associated value of a video clip is above/below five, the video clip is labeled as high/low, and similarly for the valence, if the value is above/below five, the video clip is labeled as positive/negative. Fig. 3 shows how to map different parts of the V−A space to the output labels.

3.1 Feature Extraction

Previous studies showed that low–level visual features of videos and the emotion that is evoked in their audience are correlated [11]. For instance, it has been shown that *lighting key*, *motion* and *color* have direct correlation with the affective type [18]. This has motivated our choice of these three low−level features for the affective classification of video clips.

Lighting key measures the contrast between dark and light. From the affective point of view, high−key lighting with small light/dark contrast is usually used to produce joyous scenes, whereas low−key lighting with heavy light/dark contrast is used to evoke unpleasant feelings [19]. Colours *yellow*, *orange* and *red* are related to the emotion *fea*r and *anger* while, *blue*, *violet* and *green* can evoke high valence

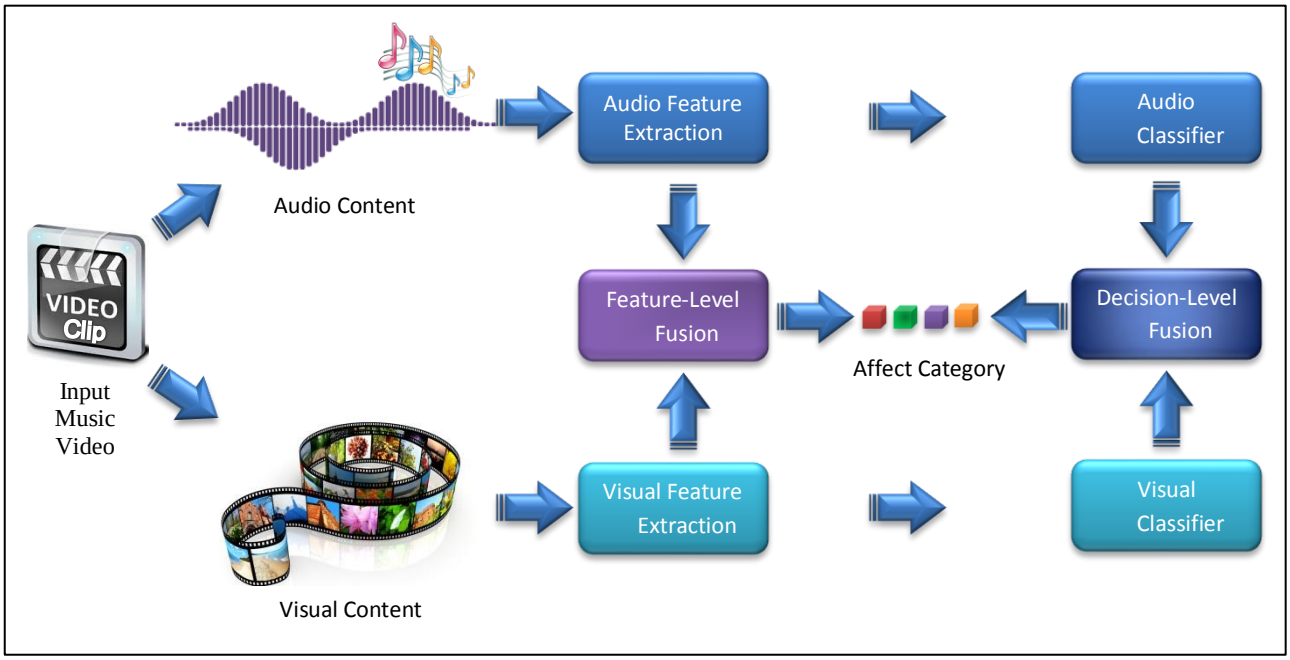

Fig. 2 an overview of the proposed system.

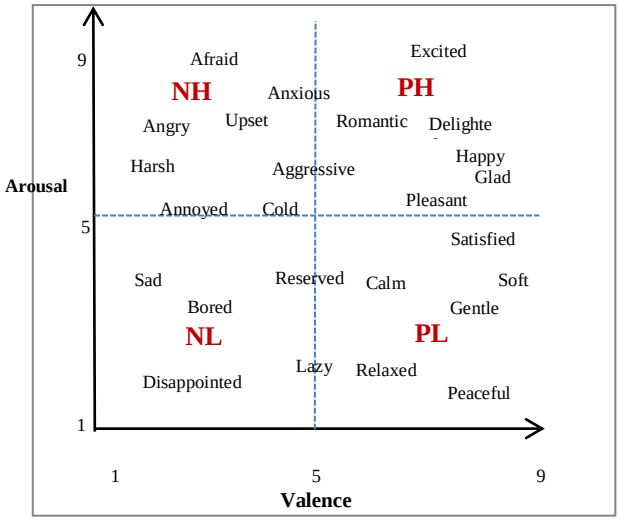

Fig. 3 Mapping V–A space quadrants to affect labels (NH, PH, NL, and PL) in the proposed system (adopted from [20]).

and low arousal emotions in the viewer [21]. There is also a relationship between emotions types *joy*, *anger*, *sadness*, and *fear* and the motion [22]. Moreover, psycho−physiological studies show that the perception of motion in a video clip is correlated with the degree of mental excitement [23].

The audio content of a video clip has also a close relationship with its affective type. Different low–level features can be extracted from the audio channel of videos to specify the affect category [11]. For example, it has been shown that arousal may be specified by considering *tempo* (fast/slow) and *pitch* features, whereas valence is better characterized with the *energy* feature [11]. In this study, we choose four popular audio features namely, *zero–crossing rate* (ZCR), *energy*, *Mel–frequency cepstral coefficients* (MFCC), and *pitch*.

3.2 Classification

As pointed out earlier, information fusion can be performed in either feature–level or decision–level. Comparative empirical studies have shown that decision–level techniques produce better results than feature−level methods. However, the choice of the fusion level is essentially based on the application [11]. In the current study, we examine both feature−level and decision–level fusion methods. Therefore, having extracted visual and audio features, they are passed to separate modules namely, *Feature−Level Fusion*, *Audio Classifier*, and *Visual Classifier* (see Fig. 2). More details about the fusion modules will be presented in the next subsection.

3.3 Fusion

For the feature−level fusion, audio and visual feature vectors are simply merged into one feature vector (Audio−Visual feature vector) and then, this feature vector is fed into a supervised classifier. It should be pointed out

that, for clarity, a separate classification module for feature−level fusion was not considered in Fig. 2. In fact the feature−level fusion module should contain a classifier as described in the previous section. Therefore, as shown in Fig. 2, the output of the feature−level fusion module is the final affect label of the input video clip.

For the decision−level fusion, having extracted visual and audio features, they are first fed into separate classifiers and then, the classification results are passed to the decision−level fusion module. Several methods were suggested for decision–level fusion including the product of confidence measures, voting, max, sum, and weighted product [24]. However, these fusion methods neither were designed for affective video retrieval, nor have any theoretical basis. To address this problem, we propose a new fusion mechanism based on the *Dempster−Shafer (DS) theory of evidence* [8] and apply it to the output of two classification modules in Fig. 2. The rationale behind the choice of the DS theory is that it is not only a well−understood formal framework for combining different sources of evidence, but also it has been successfully applied to several fusion problems in different contexts such as text categorization and sentiment analysis [8, 25, 26].

3.4 The Proposed Fusion Method

The DS theory of evidence is an effective method to quantify the degree of supports from a particular proposition based on different sources of evidence. In order to use DS theory for data fusion, the problem domain must be first identified by a finite set ϕ of mutually exclusive hypotheses, called the *frame of discernment*. The next step in applying DS theory to the fusion problem is to define a *mass function*, $m(A)$, for characterizing the strength of evidence supporting each subset $A \subseteq \phi$ based on a given piece of evidence. This function is a *basic probability assignment (BPA).* If $m(A) > 0$, the subset *A* is called a *focal element* or *focus* of *m* and if it contains only one element, *A* is called a singleton [27].

The final step for exploiting the DS theory in a fusion problem is utilizing the Dempster's rule of combination to aggregate two independent bodies of evidence (e.g. $m_{Audio} = m_A$, $m_{Visual} = m_V$) into one body of evidence as follows [8]:

$$(m_A \oplus m_V)(A) = \frac{\sum_{X \cap Y = A} m_A(X) m_V(Y)}{1 - \sum_{X \cap Y = \emptyset} m_A(X) m_V(Y)} \quad (1)$$

where the denominator is used as a normalization factor to ensure that the combination $m_A \oplus m_V$ is still a BPA.

In the current study, the outputs of audio and visual classifiers are considered as evidence for the final affect category of video clips. In the next step, we suggest the normalized probability function as follows:

$$m_d(\{c_i\}) = \frac{P(c_i|x_d)}{\sum_{j=1}^{|C|} P(c_j|x_d)} \quad (2)$$

where $m_d(\{c_i\})$ is the associated mass function for each modality d (i.e. audio and visual), $P(c_i|x_d)$ denotes the probability of a video clip belonging to class c_i given the feature vector x_d, $c \in [1, \cdots, C]$ is the final affect category to which the video clip is assigned, and C is the total number of categories (i.e. four categories in the current study). Eq. (2) may be used separately for audio and visual modalities to identify the affect category of a video clip. Finally, the overall combined decision is obtained by applying Eq. (1) to decisions made by two modalities.

In order to reduce the computational complexity of applying Eq. (1), a small partition of ϕ may be used instead of ϕ. This partition should contain as few as possible focal elements to represent possible answers to the problem [27]. However, the main difficulty of this approach is the way in which focal elements are selected. To address this problem, it is common to use only the element with the highest confidence. This approach is called DS−H in the current study.

The main problem of the DS−H method is that it may reduce the performance. This occurs when there is a dominant classifier producing high confidence values which should be always selected as the final decision [28]. To overcome this drawback, in the current study we adopt the method proposed by Bi et al [8]. Specifically, we also include the second maximum decision when combining classifiers' outputs and suggest a two−point mass function (DS−T) to partition the output of classifiers as follows:

$$m(\{f\}) + m(\{s\}) + m(\phi) = 1 \quad (3)$$

where $\{f\}$ and $\{s\}$ are focal singletons corresponding to the first and second most probable decisions and are defined as follows:

$$\{f\} = argmax\left(\left\{m(\{c_1\}), m(\{c_2\}), \ldots, m\left(\left\{c_{|\phi|}\right\}\right)\right\}\right) \quad (4)$$

$$\{s\} = argmax\left(\left\{m(\{c\}) | c \in \{c_1, c, \ldots, c_{|\phi|}\} - \{f\}\right\}\right) \quad (5)$$

Different cases may arise based on the relation between any pair of two–point mass functions [8]. Assume we are given $\langle\{f_1\},\{s_1\},\phi\rangle$ and $\langle\{f_2\},\{s_2\},\phi\rangle$:

1. Two focal points equal: this occurs when either $\{f_1\} = \{f_2\}$ and $\{s_1\} = \{s_2\}$,or $\{f_1\} = \{s_2\}$ and $\{f_2\} = \{s_1\}$. As we considered four different

affect categories in this study, the combination of such two–point mass functions contains only two different focal elements. Now, suppose for illustration, that these classes are specified by *u*, *v*, *w*, and *t*, respectively. So, the combination for two focal elements *u* and *v* is calculated as follows:

$$(m_A \oplus m_V)(\{u\}) = K[m_A(\{u\})m_V(\{u\}) + m_A(\{u\})m_V(\phi) + m_A(\phi)m_V(\{u\})] \quad (6)$$

$$(m_A \oplus m_V)(\{v\}) = K[m_A(\{v\})m_V(\{v\}) + m_A(\{v\})m_V(\phi) + m_A(\phi)m_V(\{v\})] \quad (7)$$

where *K* is again the normalization factor as follows:

$$K^{-1} = 1 - m_A(\{u\})m_V(\{v\}) - m_A(\{v\})m_V(\{u\}) \quad (8)$$

2. One focal point equal: this condition is held in one of four circumstances as follows. $\{f_1\} = \{f_2\}$ and $\{s_1\} \neq \{s_2\}$; $\{f_1\} \neq \{f_2\}$ and $\{s_1\} = \{s_2\}$; $\{f_1\} = \{s_2\}$ and $\{f_2\} \neq \{s_1\}$; $\{f_1\} \neq \{s_2\}$ and $\{f_2\} = \{s_1\}$. Here, the combination contains three different focal elements as follows (where, *u* is the common focal point):

$$(m_A \oplus m_V)(\{u\}) = K[m_A(\{u\})m_V(\{u\}) + m_A(\{u\})m_V(\phi) + m_A(\phi)m_V(\{u\})] \quad (9)$$

$$(m_A \oplus m_V)(\{v\}) = Km_A(\{v\})m_V(\phi) \quad (10)$$

$$(m_A \oplus m_V)(\{t\}) = Km_A(\phi)m_V(\{t\}) \quad (11)$$

3. Totally different focal points: In the case that all focal points are different, the combination contains four different focal elements:

$$(m_A \oplus m_V)(\{u\}) = Km_A(\{u\})m_V(\phi) \quad (12)$$

$$(m_A \oplus m_V)(\{v\}) = Km_A(\{v\})m_V(\phi) \quad (13)$$

$$(m_A \oplus m_V)(\{t\}) = Km_A(\phi)m_V(\{t\}) \quad (14)$$

$$(m_A \oplus m_V)(\{w\}) = Km_A(\phi)m_V(\{w\}) \quad (15)$$

4. Results and discussion

4.1 DEAP dataset

The DEAP (Database for Emotion Analysis using Physiological signals) dataset is a multimodal dataset for the analysis of human affective states using electroencephalogram, physiological and video signals [9]. The dataset includes ratings from an online self−assessment where 120 one−minute extracts of music videos were rated by 14−16 volunteers based on arousal, valence and dominance. Moreover, it includes participant ratings, physiological recordings and face video of an experiment where 32 volunteers watched a subset of 40 of the music videos. EEG and physiological signals were recorded and each participant also rated the videos as above [9]. In the current study, we have used all the video clips whose YouTube links are provided in the DEAP dataset and that were available on YouTube at the time when experiments were conducted (totally 43 music clips). In the DEAP dataset, arousal and valence values are identified with integers in the range of 1 to 9. These numbers correspond to the range *calm/bored* to *stimulated/excited* for the arousal values, whereas they associate with the range *unhappy/sad* to *happy/joyful* for the valence values. However, as pointed out earlier, in the current study we aim at classifying video clips into four affect categories (i.e. NH, NL, PH, and PL) each corresponding to one quadrant of the V−A space. Therefore, we used the online ratings of video clips provided within the DEAP dataset as follows. First, the average of ratings is computed for the arousal and valence values, respectively. Then, if values are above/below five, the video clip is labeled as high/low and positive/negative for the arousal and valence dimensions, respectively (see Fig. 3 for more details).

4.2 Experimental Setup

As described in previous section, different features were extracted from the audio–visual contents of 43 music video clips. Specifically, from the audio channel of video clips we extracted the 13−dimensional MFCC features, ZCR, energy, and pitch. Similarly, visual features are extracted as follows. For each frame, the 16 bin color histograms for the hue component of the HSV space are calculated. Then, for specifying the lighting key, after computing the 16 bin histograms of the value component of the HSV space, the mean and the standard deviation of the histogram are multiplied. Moreover, a motion vector is computed for every four frames in video sequences and a block size of 16 is used in a block matching algorithm [29]. Having calculated the motion vector of each frame, the mean of their absolute values are added to previously extracted low–level visual features. Two popular supervised machine learning algorithms are used for classification namely, SVM and Naïve Bayes. It should be noted that we performed both feature−level and decision−level fusion. Hence, for feature−level fusion, the associated feature vector of each modality is first merged into one feature vector and then, the combined vector is classified. For the decision−level, on the other hand, audio and visual feature vectors are independently used to classify a video clip and then, classification results are combined using fusion methods.

In order to evaluate the proposed method, the *precision*, *recall*, and *F–measure* are used in the experiments. These measures are frequently used in machine learning and information retrieval researches and are computed as follows [13]:

$$\mathrm{F-measure} = \frac{2\times(precision+recall)}{precision+recall} \tag{16}$$

where precision and recall defined as follows:

$$\mathrm{precision} = \frac{TP}{TP+FP} \tag{17}$$

$$\mathrm{recall} = \frac{TP}{TP+FN} \tag{18}$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively.

4.3 The Effectiveness of Proposed Fusion Method

In order to assess the utility of using the proposed fusion method for affective video retrieval, we first performed our experiments on the audio and visual modalities independently. Then, we investigated the fusion of audio−visual modalities. The precision and recall of the proposed SVM and Naïve Bayes are showed in Table 1 for six fusion methods namely, feature−level fusion (FL), product rule, maximum, sum, traditional Dempster–Shafer rule (DS−H), and two−point Dempster–Shafer rule (DS−T). In our experiments we aim at addressing the following research questions:

1. Can the fusion of audio–visual contents improve the performance of affective video retrieval?
2. Perfuming the fusion at which level is more suitable for multimodal affective video retrieval?
3. Can applying DS theory of evidence to multimodal data fusion increase the performance of affective video retrieval system?
4. What is the effect of using a two−point mass function which also considers the second maximum decision when combining classifiers' outputs?

Similarly, The F–Measure of SVM and Naïve Bayes classifiers are depicted in Fig. 4, and Fig. 5 for the above mentioned fusion methods. As can be seen in Table 1, Fig. 4 and Fig. 5, the classification accuracy of unimodal approach depend on the classification method. Specifically, for the Naïve Bayes algorithm, the performance of using only visual features is much higher than the performance of exploiting only audio features. In contrast, for the SVM classifier, the performance of audio modality is slightly better than the performance of visual modality. However, for both algorithms, four out of five decision−level fusion methods perform better than unimodal approach. Going back to our first research question, these results show that the fusion of audio–video contents significantly improves the performance of affective video retrieval systems.

Another notable result in Table 1, Fig. 4, and Fig. 5 is that regardless of which classification algorithm is used, the performance of all decision−level fusion methods are higher than the performance of feature−level fusion method. Therefore, the answer to the second research question would be "the decision−level fusion is more suitable for multimodal affective video retrieval".

In order to address the third research question, the performance of two DS−based fusion methods should be compared with three mentioned decision−level fusion methods. As shown in Table 1, Fig. 4, and Fig. 5, using the SVM algorithm, both DS−based methods outperform other approaches whereas, using the Naïve Bayes algorithm, only the performance of the DS−T method is higher than other fusion methods. Therefore, the third research question is also answered: the DS−based fusion method can be successfully applied to the affective video retrieval problem. Moreover, the forth question may be answered as follows. In order to obtain the best performance using the DS theory of evidence, it is necessary to consider the

second maximum decision when combining classifiers' outputs.

Table1. Comparison of precision and recall for SVM and Naïve Bayes Classifiers.

		Classifier	SVM		Naïve Bayes	
		Measures	Precision	Recall	Precision	Recall
		Audio	0.6519	0.6512	0.5374	0.5116
		Visual	0.6475	0.6279	0.7628	0.6977
Fusion Methods	Audio + Visual	FL	0.634	0.635	0.6369	0.6047
		Product	0.6907	0.6744	0.7458	0.6279
		Max	0.6698	0.6744	0.7611	0.6977
		Sum	0.6452	0.6512	0.7697	0.6744
		DS−H	0.7417	0.6744	0.7446	0.6279
		DS−T	**0.7563**	**0.6744**	**0.7818**	**0.7209**

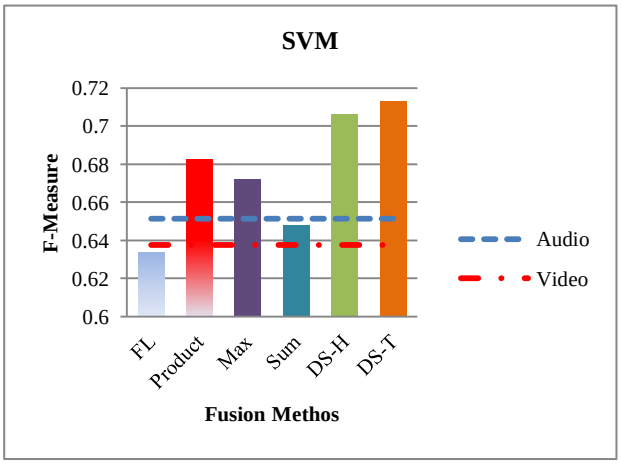

Fig. 4 Comparison of the F–Measure of classifying video clips using feature−level fusion (FL), product, max, sum, Dempster (DS−H), and proposed method (DS−T) for the SVM classifier.

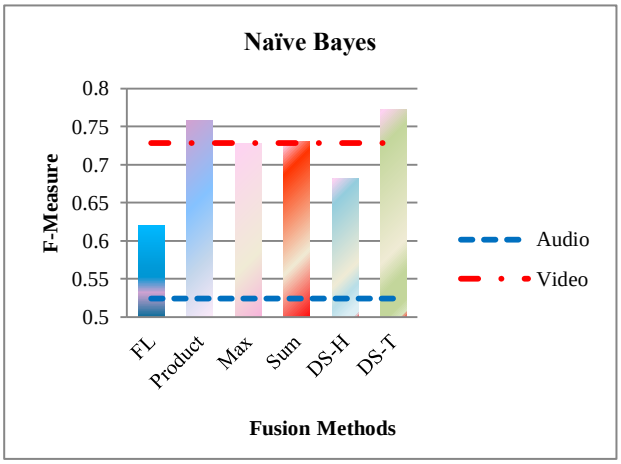

Fig. 5 Comparison of the F–Measure of classifying video clips using feature−level fusion (FL), product, max, sum, Dempster (DS−H), and proposed method (DS−T) for the Naïve Bayes classifier.

5. Conclusions and Future Work

In this study a multimodal approach was proposed for affective video retrieval and the impact of combining the audio–visual contents of video clips on the performance of the affective video retrieval system was investigated. The goal of the proposed system is to classify each music video clip into one of the four quadrants of the V–A space (valence–arousal space). In order to achieve this goal, low–level audio–visual features are used and both decision−level and feature–level fusion is performed. In the feature–level fusion, having extracted audio−visual features, they are first merged into one feature vector, and then, fed into the classification method. In the decision–level, on the other hand, video clips are classified based on audio and visual modalities independently, and then the results is combined to specify the overall affect category of the video clip. For the fusion step of the proposed system, two evidential approaches based on the Dempster–Shafer theory of evidence are suggested, namely DS–H and DS–T methods. In the DS−H method, only the decision with the highest confidence is used in the combination, whereas the DS–T method also includes the second maximum decision when combining classifiers' outputs. We compared these methods with three frequently used fusion methods namely, the product rule, sum rule, and the maximum method. Experiments were conducted on the DEAP dataset. Two state–of–the–art supervised machine learning algorithms namely, SVM and Naïve Bayes are used for classification. Experimental results indicate that feature–level fusion does not provide suitable results in comparison to decision–level approaches. Furthermore, results show that the proposed DS–T method outperforms the other fusion methods.

The main application of the proposed system is in video recommendation systems. In particular, the proposed system helps video delivery websites to provide more accurate and convincing video recommendations by efficiently considering the affective content of movies. Also, with the aim of the proposed system videos can be effectively produced to enhance the intended emotion of viewers.

The main contributions of this work are as follows: improving the performance of affective video retrieval systems by adapting a fusion method based on the Dempster−Shafer theory of evidence; Investigating the effect of applying different fusion methods for affective video retrieval; Comparing the performance of fusion methods in both feature–level and decision–level.

The proposed system for affective video retrieval uses low−level audio–visual features, because extracting such features is more computationally efficient. It seems that the proposed system may be improved by incorporating high−level features. This can be considered as future research. Another line of future research will be investigating other mathematical theories for fusion (e.g. Bayesian data fusion). Finally, in order to develop more efficient affective video retrieval systems, another modality can be considered to complement audio–visual modalities. This can be also considered as another direction for future work.

References

[1] L. Zhaoming, W. Xiangming, L. Xinqi, and Z. Wei, "A Video Retrieval Algorithm Based On Affective Features," in *Computer and Information Technology, 2009. CIT'09. Ninth IEEE International Conference on*, 2009, pp. 134−138.

[2] A. Yazdani, K. Kappeler, and T. Ebrahimi, "Affective content analysis of music video clips," in *Proceedings of the 1st international ACM workshop on Music information retrieval with user−centered and multimodal strategies*, 2011, pp. 7−12.

[3] R. M. A. Teixeira, T. Yamasaki, and K. Aizawa, "Determination of emotional content of video clips by low−level audiovisual features," *Multimedia Tools and Applications*, vol. 61, pp. 21−49, 2012.

[4] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *Affective Computing, IEEE Transactions on*, vol. 1, pp. 18−37, 2010.

[5] M. S. Hussain, R. A. Calvo, and P. A. Pour, "Hybrid fusion approach for detecting affects from multichannel physiology," in *Affective computing and intelligent interaction*, ed: Springer, 2011, pp. 568−577.

[6] D. Datcu and L. J. Rothkrantz, "Emotion recognition using bimodal data fusion," in *Proceedings of the 12th International Conference on Computer Systems and Technologies*, 2011, pp. 122−128.

[7] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding Affective Content of Music Videos through Learned Representations," in *MultiMedia Modeling*, 2014, pp. 303−314.

[8] Y. Bi, "The impact of diversity on the accuracy of evidential classifier ensembles," *International Journal of Approximate Reasoning*, vol. 53, pp. 584−607, 2012.

[9] S. Koelstra, C. Muhl, M. Soleymani, J.−S. Lee, A. Yazdani, T. Ebrahimi, *et al.*, "Deap: A database for emotion analysis; using physiological signals," *Affective Computing, IEEE Transactions on*, vol. 3, pp. 18−31, 2012.

[10] G. Valenza and E. P. Scilingo, "Emotions and Mood States: Modeling, Elicitation, and Classification," in *Autonomic Nervous System Dynamics for Mood and Emotional−State Recognition*, ed: Springer, 2014, pp. 9−21.

[11] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi, "Multimedia content analysis for emotional characterization of music video clips," *EURASIP Journal on Image and Video Processing*, vol. 2013, pp. 1−10, 2013.

[12] A. Hanjalic and L.−Q. Xu, "Affective video content representation and modeling," *Multimedia, IEEE Transactions on*, vol. 7, pp. 143−154, 2005.

[13] S. Wang, Y. Zhu, G. Wu, and Q. Ji, "Hybrid video emotional tagging using users' EEG and video content," *Multimedia Tools and Applications*, pp. 1−27, 2013.

[14] G. Caridakis, K. Karpouzis, and S. Kollias, "User and context adaptive neural networks for emotion recognition," *Neurocomputing*, vol. 71, pp. 2553−2562, 2008.

[15] K. R. Scherer and H. Ellgring, "Multimodal expression of emotion: Affect programs or componential appraisal patterns?," *Emotion*, vol. 7, p. 158, 2007.

[16] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and emotion in human−computer interaction*, ed: Springer, 2008, pp. 92−103.

[17] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, pp. 39−58, 2009.

[18] C. Huang, T. Fu, and H. Chen, "Text‐based video content classification for online video‐sharing sites," *Journal of the American Society for Information Science and Technology*, vol. 61, pp. 891−906, 2010.

[19] M. Xu, C. Xu, X. He, J. S. Jin, S. Luo, and Y. Rui, "Hierarchical affective content analysis in arousal and valence dimensions," *Signal Processing*, vol. 93, pp. 2140−2150, 2013.

[20] J.−Y. Liu, S.−Y. Liu, and Y.−H. Yang, "LJ2M DATASET: TOWARD BETTER UNDERSTANDING OF MUSIC LISTENING BEHAVIOR AND USER MOOD."

[21] H.−B. Kang, "Affective content detection using HMMs," in *Proceedings of the eleventh ACM international conference on Multimedia*, 2003, pp. 259−262.

[22] K. Sun and J. Yu, "Video affective content representation and recognition using video affective tree and hidden markov models," in *Affective Computing and Intelligent Interaction*, ed: Springer, 2007, pp. 594−605.

[23] H. L. Wang and L.−F. Cheong, "Affective understanding in film," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, pp. 689−704, 2006.

[24] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *Affective Computing, IEEE Transactions on*, vol. 3, pp. 211−223, 2012.

[25] M. E. Basiri, N. Ghasem−Aghaee, and A. R. Naghsh−Nilchi, "Exploiting reviewers' comment histories for sentiment analysis," *Journal of Information Science*, vol. 40, pp. 313−328, 2014.

[26] M. E. Basiri, A. R. Naghsh−Nilchi, and N. Ghasem−Aghaee, "Sentiment Prediction Based on

Dempster−Shafer Theory of Evidence," *Mathematical Problems in Engineering,* vol. 2014, 2014.

[27] D. A. Bell, J.−w. W. Guan, and Y. Bi, "On combining classifier mass functions for text categorization," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 17, pp. 1307−1319, 2005.

[28] Y. Bi, J. Guan, and D. Bell, "The combination of multiple classifiers using an evidential reasoning approach," *Artificial Intelligence*, vol. 172, pp. 1731−1751, 2008.

[29] A. Barjatya, "Block matching algorithms for motion estimation," *IEEE Transactions Evolution Computation,* vol. 8, pp. 225−239, 2004.

Shahla Nemati received the B.Eng. degree in hardware computer engineering in 2004 at the Shiraz University in Iran. She also received the M.Sc. degree in 2007 at the Isfahan University of Technology (IUT), Iran. She is currently a PhD candidate of the Department of Computer Architecture, University of Isfahan, Iran. Her current research interests include video processing, music/audio processing and affective computing.

Ahmad Reza Naghsh−Nilchi received the B.E., M.E., and Ph.D. degrees in electrical engineering from the University of Utah, United States, in 1990, 1991, and 1997, respectively. He is currently an Associate Professor at the Department of Artificial Intelligent, University of Isfahan, Iran. His current research interests include medical Signal and image Processing, Digital Data Hiding (Watermarking, Steganography, Encryption, etc.), Computer Graphics, Sinc-Convolution. Dr. Naghsh Nilchi is currently the Chairman and Faculty member of the Department of Artificial Intelligence and Multimedia at the University of Isfahan, Iran and the Editor−in−Chief of Journal of Computing and Security (JCS).