Fusion of Learned Multi-Modal Representations and Dense Trajectories for Emotional Analysis in Videos

Esra Acar
DAI Laboratory
Technische Universität Berlin
Berlin, Germany
esra.acar@tu-berlin.de

Frank Hopfgartner
Humanities Advanced Technology and Information Institute
University of Glasgow
Glasgow, UK
frank.hopfgartner@glasgow.ac.uk

Sahin Albayrak
DAI Laboratory
Technische Universität Berlin
Berlin, Germany
sahin.albayrak@dai-labor.de

***Abstract*—When designing a video affective content analysis algorithm, one of the most important steps is the selection of discriminative features for the effective representation of video segments. The majority of existing affective content analysis methods either use low-level audio-visual features or generate handcrafted higher level representations based on these low-level features. We propose in this work to use deep learning methods, in particular convolutional neural networks (CNNs), in order to automatically learn and extract mid-level representations from raw data. To this end, we exploit the audio and visual modality of videos by employing Mel-Frequency Cepstral Coefficients (MFCC) and color values in the HSV color space. We also incorporate dense trajectory based motion features in order to further enhance the performance of the analysis. By means of multi-class support vector machines (SVMs) and fusion mechanisms, music video clips are classified into one of four affective categories representing the four quadrants of the Valence-Arousal (VA) space. Results obtained on a subset of the DEAP dataset show (1) that higher level representations performs better than low-level features, and (2) that incorporating motion information leads to a notable performance gain, independently from the chosen representation.**

I. Introduction

Nowadays, the amount of audio-visual data items available to consumers has attained colossal proportions. Delivering personalized video content corresponding to the needs of the consumers is a challenge which still has to be resolved. Video affective analysis can bring an answer to such a challenge from an original perspective. In particular, in the context of categorical affective analysis, where human emotions are defined in terms of discrete categories (opposed to dimensional affective analysis where they are non-discrete [1]), one direction followed by many researchers consists in using machine learning methods (e.g., [2]–[4]). Machine learning approaches make use of a specific data representation (i.e., features extracted from the data) to identify particular events. However, their performance is heavily dependent on the choice of the data representation on which they are applied [5]. As in any pattern recognition task, one key issue is, therefore, to find an effective representation of video content.

Features can be classified according to different schemes. One type is the classification based on the level of semantic information which a given feature carries. In the terminology which we adopt, at one extreme, a feature is said to be low-level if it carries (almost) no semantic information (e.g., value of a single pixel or audio sample); at the other extreme, it is said to be high-level if it carries maximally semantic information (e.g., a guitarist performing a song in a clip). Between both, mid-level feature representations are derived from raw data, but are one step closer to human perception. Another possible type of classification, which is particularly relevant in video analysis, where data items are not a single image but sequences, is the distinction between static and dynamic (or temporal) features.

In this context, in the field of audio, video and more generally multi-dimensional signal processing, automatically and directly learning suitable features (i.e., mid-level features) from raw data to perform tasks such as event detection, summarization, retrieval has attracted particular attention, especially because such learning kept the amount of required supervision to a minimum and provided scalable solutions. To achieve this, deep learning methods such as CNNs and deep belief networks are shown to provide promising results (e.g., [6]–[8]). This recent success of deep learning methods previously incited us to directly learn feature representations from automatically extracted raw audio and color features by deep learning to obtain mid-level audio and visual representations [9]. However, this work was limited to learning audio and static visual features only. As a matter of fact, our work could not optimally take into account the motion and temporal coherence exhibited in image sequences compared to a single image. Studies (e.g., [3], [10]) have, indeed, demonstrated that, in addition to audio and static color features, motion plays an important role for affective content analysis in edited videos such as movies and music video clips. Therefore, we propose to use dense trajectory features to derive a mid-level motion representation obtained via a sparse coding based Bag-of-Words method to boost the classification performance of our system. In our previous work, static color features were based on the representation of images in the RGB space, which is not optimized for affect analysis. Hinted by the prior art evidence that perception of emotions by humans is enhanced in a hue-saturation type color space [11], we close this gap here by working in the HSV color space. One additional question arising when using multiple features, is that of fusion of information. We also propose here an assessment

978-1-4673-6870-4/15/$31.00

of fusion mechanisms. Consequently, the aim of this work is to (1) investigate the discriminative power of learned audio and static visual representations in the HSV space, (2) assess the effect of incorporating dense trajectories, and (3) investigate optimal fusion mechanisms for combining audio, static visual and dynamic visual features. To the best of our knowledge, the proposed system is the first to adopt dense trajectories as motion features for emotional characterization of music videos; and we show that the mid-level motion and learned audio-visual representations are more discriminative and provide more accurate results than low-level audio-visual ones.

The paper is organized as follows. Section II explores the recent developments and reviews methods which have been proposed for affective content analysis of video material with an emphasis on the feature representation of videos. In Section III, we introduce our method for the affective classification of music video clips. We provide and discuss evaluation results on a subset of the DEAP dataset [12] in Section IV. Finally, we present concluding remarks and future directions to expand our method in Section V.

II. Related Work

Among video affective content analysis methods, using low-level audio-visual features as video representations is one type of commonplace approach. In [13], a method for mood-based classification of TV Programs on a large-scale dataset is presented, in which frame-level audio-visual features are used as video representations. In [14], a combined analysis of low-level audio and visual representations based on early feature fusion is presented for facial emotion recognition in videos. The baseline framework introduced in [15] also employs low-level audio and still image features.

Another type of commonplace approach is to use mid-level or hierarchical representations of videos. These solutions employ mid-level representations created from low-level ones. Irie et al. [16] present an affective video segment retrieval method based on the correlation between emotions and so-called emotional audio events (EAEs) which are *laughter*, *loud voice*, *calm music* and *aggressive music*. The main idea is to use EAEs as an intermediate representation. Xu et al. [4] present a 3-level affective content analysis framework, in which the purpose is to detect the affective content of videos (i.e., horror scenes for horror movies, laughable sections for sitcoms and emotional tagging of movies). They introduce mid-level representations which indicate dialog, audio emotional events (i.e., horror sound and laughter) and textual concepts (i.e., informative emotion keywords). In [2], Irie et al. propose to represent movie shots with so-called Bag-of-Affective Audio-visual Words and apply a latent topic driving model in order to map these representations to affective categories. In [17], Canini et al. introduce a framework where movie scenes are represented in a 3-dimensional connotative space whose dimensions are *natural*, *temporal*, and *energetic*. The aim is to reduce the gap between objective low-level audio-visual features and highly subjective emotions through connotation. As audio-visual representation of movies, they employ low-level audio descriptors, low and mid-level color and motion descriptors. In [18], Jiang et al. propose a comprehensive computational framework, where they extract an extensive set of features from the dataset, ranging from well-known low-level audio-visual descriptors to high level semantic attributes such as ObjectBank and SentiBank representations.

All of the abovementioned works represent videos with low or mid-level handcrafted features. However, in attempts to extend the applicability of methods, there is a growing interest for directly and automatically learning features from raw audio-visual data rather than representing them based on manually designed features. For example, Schmidt et al. [8] address the feature representation issue for automatic detection of emotions in music by employing regression based deep belief networks to learn features from magnitude spectra instead of manually designing feature representations. By taking into account the dynamic nature of music, they also investigate the effect of combining multiple timescales of aggregated magnitude spectra as a basis for feature learning. These learned features are then evaluated in the context of multiple linear regression. Li et al. [7] propose to perform feature learning for music genre classification and use CNNs for the extraction of musical pattern features. Ji et al. [6] address the automated recognition of human actions in surveillance videos and develop a novel 3D-CNN model to capture motion information encoded in multiple adjacent frames. Another CNN-based method is our previous work [9], where we used deep learning to derive mid-level representations directly from the raw data.

One observation about the works mentioned above is that the use of the temporal aspect of videos is either limited or totally absent. In other words, videos are generally analyzed as a sequence of independent frames rather than a whole. A few works use motion-based features, and these are limited to simple features (e.g., features based on frame differencing). The only notable exception is the work of Ji et al. [6], where multiple adjacent frames are used. However, they take into account only 7 adjacent frames. Increasing this number to higher dimensions would probably render the learning of the 3D-CNNs intractable. Therefore, in our opinion, a more effective mid-level motion representation is needed.

Recently, a new type of video descriptor has emerged, namely dense feature trajectories. These descriptors, which correspond to points which are densely sampled and tracked using dense optical flow fields, were introduced by Wang et al. [19] for the task of action recognition in videos, and have proven robust for action recognition. However, to the best of our knowledge, the applicability of these dense trajectories to the task of affective content analysis has not been investigated yet. Distinct from the aforementioned existing works, we suggest combining deep learning based representations with dense motion trajectories. In other words, we propose to learn both audio and static visual feature representations by using a CNN and perform the affective classification of music video clips by fusing these representations with dense trajectory based motion features at the decision-level. We also show that the mid-level motion and learned audio-visual representations are more discriminative than low-level audio-visual ones.

III. The Video Affective Analysis Method

In this section, we present our approach, which is a categorical affective analysis solution. It performs classification of music video clips into one of the four quadrants of the

Valence-Arousal-space (VA). As mentioned in the introduction, affective analysis can either be categorical or dimensional. The choice of categorical or dimensional is not critical, as in practice, categories can always be mapped onto dimensions and vice versa [1]. It is, therefore, possible to map discrete emotions to arousal-valence dimensions.

The system consists of the following steps: (1) one-minute highlight extracts of music video clips are first segmented into pieces, each piece lasting 5 seconds (as suggested in [3]); (2) audio and visual feature extraction; (3) learning mid-level audio and static visual representations (training phase only); (4) generating mid-level audio-visual representations; (5) generating an affective analysis model (training phase only); (6) classifying a video segment of 5-second length into one of the four quadrants in the VA-space (test phase only); and (7) classifying a complete music video clip using the results obtained on the 5-second segments constituting the clip (test phase only).

The audio and visual feature learning phases are discussed in detail in Section III-A, whereas the incorporation of temporal information to the system is explained in Section III-B. The generation of an affective analysis model is discussed in more detail in Section III-C. This model uses fusion, which is presented in Section III-D.

A. Learning Mid-Level Audio and Static Visual Representations

Concerning the learning of audio and visual representations, we improved one of our previous works on affective content analysis [9]. The improvements concern the extraction modalities (e.g., dimensions) of the audio representations, and the use of a different color space which enables deriving more discriminative features. MFCC values are extracted for each video segment. The resulting MFCC feature vectors are given as input to a CNN. The first layer (i.e., the input layer) of the CNN is a 497x13 map which contains the MFCC feature vectors from 125 frames of one music video segment. In Figure 1, the CNN architecture used to generate audio representations is presented. The CNN has three convolution and two subsampling layers, and one output layer which is fully connected to the last convolution layer (this network size in terms of convolution and subsampling layers has experimentally given satisfactory results). The output layer consists of four units: one for each quadrant of the VA-space, where each unit is fully connected to each of the 976 units in the last convolution layer. The CNN is trained using the backpropagation algorithm. After training, the output of the last convolution layer is used as the mid-level audio representation of the corresponding video segment. Hence, the MFCC feature vectors from 125 frames of one segment are converted into a 976-dimensional feature vector (which constitutes a more abstract audio representation) capturing the acoustic information in the audio signal of the music video segment.

Existing works (e.g., [11]) have shown that colors and their proportions are important parameters to evoke emotions. This observation has motivated our choice of color values for the generation of static visual representations for music videos. The frame in the middle of a 5-second video segment is extracted as the keyframe (i.e., representative frame) for the segment. For the generation of mid-level static visual representations, we extract color information in the HSV space from the keyframe. The resulting values in each channel are given as input to a separate CNN. Similarly to the audio case, Figure 1 presents the CNN architecture used to generate visual representations, where the first layer (i.e., the input layer) of the CNN is a 160x120 map which contains the values from one channel of the keyframe. The training of the CNN is done similarly to the training of the CNN in the audio case. As a result, the values in each channel are converted into an 88-dimensional feature vector. The feature vectors generated for each of the three channels are concatenated into a 264-dimensional feature vector which forms a more abstract visual representation capturing the color information in the keyframe of the segment.

B. Deriving Mid-Level Dynamic Visual Representations

The importance of motion in edited videos such as movies and music video clips motivated us to extend our previous approach [9] and to incorporate motion information to our analysis framework. To this end, we adopt the work of Wang et al. on dense trajectories [19]. Dense trajectories are dynamic visual features which are derived from tracking densely sampled feature points in multiple spatial scales. Although initially used for unconstrained video action recognition [19], dense trajectories constitute a powerful tool for motion or video description, and, hence, are not limited to action recognition only.

Our dynamic visual representation works as follows. First, dense trajectories [19] of length 15 frames are extracted from each video segment. The sampling stride, which corresponds to the distance by which extracted feature points are spaced, is set to 20 pixels. Dense trajectories are subsequently represented by a histogram of oriented gradients (HoG), a histogram of optical flow (HoF) and motion boundary histograms in the x and y directions (MBHx and MBHy, respectively). We learn a separate dictionary for each dense trajectory descriptor (i.e., each one of HoG, HoF, MBHx and MBHy). We employ the sparse dictionary learning technique presented in [20]. In order to learn the dictionary of size k ($k = 512$ in this work) for sparse coding, $400 \times k$ feature vectors are sampled from the training data (this figure has experimentally given satisfactory results). In the coding phase, we construct the sparse representations of dense trajectory features using the LARS algorithm [21]. Given dense trajectory features and a dictionary as input, the LARS algorithm returns sparse representations for the feature vectors (i.e., sparse mid-level motion representations). In order to generate the final sparse representation of video segments which are a set of dense trajectory feature vectors, we apply the *max-pooling* technique.

C. Generating the Affective Analysis Model

In order to generate affective analysis models, mid-level audio, dynamic and static visual representations are fed into separate multi-class SVMs. In the test phase, mid-level audio and static visual representations are created by using the corresponding CNN models for music video segments of 5-second length. The music video segments are then classified

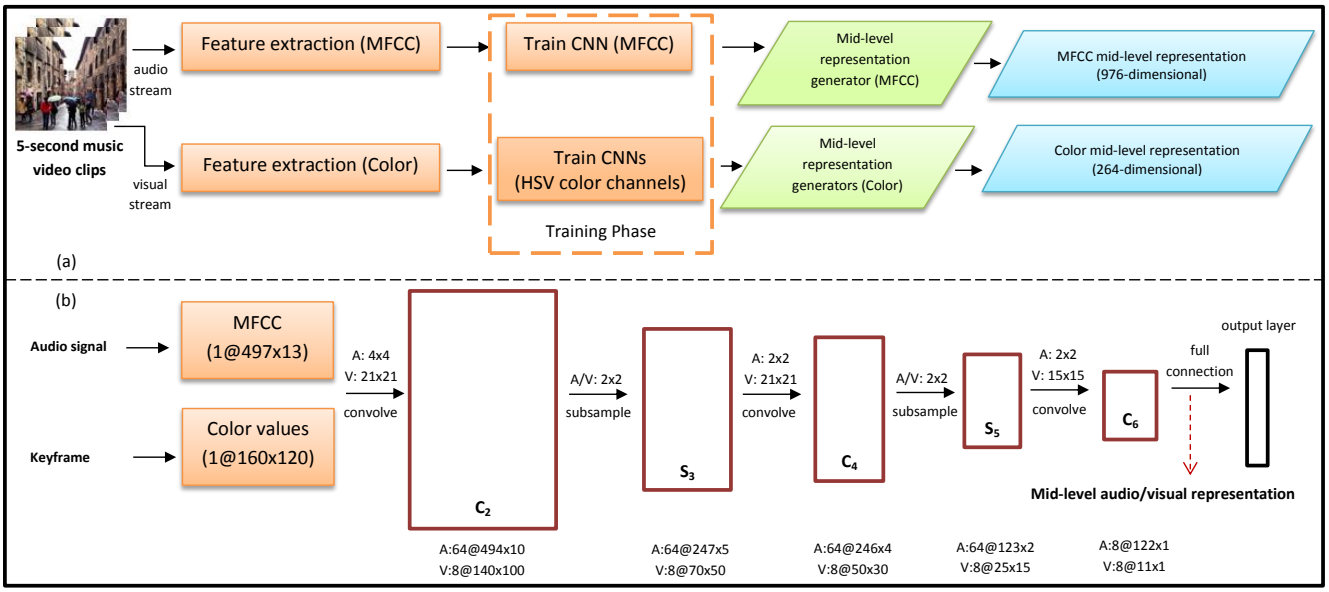


Fig. 1. (a) A high-level overview of our representation learning method, (b) the detailed CNN architectures for audio and visual representation learning. The architecture contains three convolution and two subsampling layers, one output layer fully connected to the last convolution layer, C_6. (CNN: Convolutional Neural Network, MFCC: Mel-Frequency Cepstral Coefficients, A: Audio, V: Visual)

using the three affective video analysis models (i.e., one model for each of audio, static visual and dynamic visual features).

The probability estimates of the outputs of the models are merged using one of the fusion strategies presented in Section III-D. Normally, in a basic SVM, only class labels or scores are output. The class labels result from thresholding the scores output by the SVM, which are not a probability measure. In order to enable a fusion in a probabilistic fashion, the scores returned by the SVM are converted into probability estimates using the method explained in [22].

Once all 5-second video segments extracted from a given clip are classified, final decisions for the classification of the complete music video clip is realized by a *plurality voting* process. In other words, a music video clip is assigned the label which is most frequently encountered among the set of 5-second segments constituting the clip.

D. Fusion strategies

When combining results of multiple classifiers, fusion of the results constitutes an important step. In this paper, we investigate two distinct fusion techniques to combine the outputs of the SVM models, namely *linear fusion* and *SVM-based fusion*.

1) Linear fusion*:* In linear fusion, probability estimates obtained from the SVMs trained separately with one of the mid-level audio, static visual and dynamic visual representations are fused at the decision-level using different weights for each modality. The weights are optimized on the training data.

2) SVM-based fusion*:* In SVM-based fusion, probability estimates of the SVMs are concatenated into vectors and used to construct higher level representations for each video segment. Another SVM classifier which takes as input these higher level representations is constructed. This SVM is subsequently used to predict the label of a video segment.

IV. Performance Evaluation

The experiments presented in this section aim at comparing the discriminative power of our method which is based on mid-level dense trajectory based motion and learned audio-visual representations against the method that uses low-level audio-visual features (i.e., the baseline method), and the works presented in [3] and [9]. An overview of the DEAP dataset is provided in Section IV-A. In Section IV-B, we present the experimental setup. Finally, we provide results and discussions in Section IV-C.

A. Dataset and Ground-truth

The DEAP dataset is a dataset for the analysis of human affective states using electroencephalogram, physiological and video signals. It consists of the ratings from an online self-assessment where 120 one-minute extracts of music videos were each rated by 14-16 volunteers based on arousal, valence and dominance. We have used all the music video clips whose YouTube links are provided in the DEAP dataset and that were available on YouTube at the time when experiments were conducted (74 music clips). Only one-minute highlight extracts from these 74 videos have been used in the experiments. The extracts of different affective categories downloaded from YouTube equate to 888 music video segments of 5-second length.

We have four affective labels used for classification. These are *high arousal-high valence (ha-hv)*, *low arousal-high valence (la-hv)*, *low arousal-low valence (la-lv)* and *high arousal-low valence (ha-lv)* each representing one quadrant in the VA-space. The labels are provided in the dataset and are determined by the average ratings of the participants in the online self-assessment. In our experiments, we had 22 songs of category *ha-lv*, 19 songs of category *ha-hv* and *la-hv*, and 14 songs of category *la-lv*.

B. Experimental Setup

The MIR Toolbox v1.4[1] is employed to extract the 13-dimensional MFCC features. Frame sizes of 25 ms with 10 ms overlap are used. Mean and standard deviation for each dimension of the MFCC feature vectors are computed, which compose the low-level audio representations (LLR audio) of music video segments. In order to generate the low-level visual features (LLR visual) of music video segments, we constructed normalized HSV histograms (16, 4, 4 bins) in the HSV color space. The Deep Learning toolbox[2] is used in order to generate mid-level audio and static visual representations with a CNN. Wang's implementation[3] is used to extract dense trajectories from video segments.

Computationally, the most expensive phase of the representation learning is the training of the CNNs which takes on average 150 and 350 seconds per epoch for MFCC and color features, respectively. The generation of feature representations using CNNs amounts to 0.5 and 1.2 seconds on average per 5-second video segment for MFCC and color features, respectively. The extraction of dense trajectories takes on average 16 seconds per 5-second video segment. All the timing evaluations were performed with a machine with 2.40GHz CPU and 8GB RAM.

The multi-class SVMs with an RBF kernel are trained using libsvm[4] as the SVM implementation. Training was performed using audio and visual features extracted at the music video segment level. More specifically, we trained one SVM using the CNN based mid-level audio features (MLR audio), one SVM using the CNN based mid-level static visual features (MLR static visual), a third SVM using Bag-of-Words representations based on the motion features (i.e., HoG, HoF, MBHx and MBHy) of dense trajectories (MLR motion) and another two SVMs using the LLR audio and LLR visual features as input, respectively. Due to the small amount of music video samples, we used the leave-one-song-out cross validation scheme. The SVM parameters were optimized by 5-fold cross-validation on training sets. Fusion of audio and visual features is performed at the decision-level by linear or SVM-based fusion, as explained in Section III-D.

C. Results and Discussions

First, we present the classification accuracies in the case where only one type of descriptor is employed in Table I. This gives an estimation of the influence of each descriptor (audio, static visual or dynamic visual) on the performance in detail.

TABLE I. CLASSIFICATION ACCURACIES ON THE DEAP DATASET (WITH UNIMODAL REPRESENTATIONS: AUDIO OR VISUAL-ONLY)

Method	Accuracy (%)
Our method (MLR motion)	**51.35**
Our method (MLR audio)	48.65
Our method (MLR static visual)	43.24
The LLR audio based method	37.84
The LLR visual based method	28.38

[1] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

[2] https://github.com/rasmusbergpalm/DeepLearnToolbox/

[3] http://lear.inrialpes.fr/people/wang/dense_trajectories

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

One significant point which can be inferred from Table I is that motion and audio representations are more discriminative than static visual features. This is true for the mid-level motion and learned audio representations outperforming static visual ones, and also for the low-level audio representations compared to the low-level static visual ones. The superiority of the dynamic visual feature (dense motion trajectories) can be explained by the fact that affect present in video clips is often characterized by motion (e.g., camera motion). Another important point is that learning mid-level representations consistently yields better results than low-level ones; a similar conclusion was drawn in [9]. Another important observation is the performance gain (around 15%) of using learned static visual features compared to low-level ones. When evaluated together with our previous findings about learning color representations in [9], we can conclude that color values in the HSV space lead to more discriminative mid-level representations than color values in the RGB space.

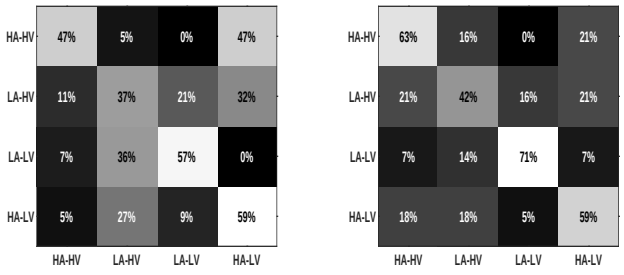
Second, we provide the classification accuracies of our method which employs MLR audio, motion and static visual representations compared to the baseline method and the works [3] and [9] on the DEAP dataset (Table II). Our method outperformed the baseline method, [3] and [9], by achieving 58.11% and 66.22% accuracy for linear and SVM-based fusion, respectively. The performance gain over prior works is particularly remarkable for SVM-based fusion, which shows that an advanced fusion mechanism can better emphasize classification performance compared to linear fusion.

These results demonstrate the potential of our approach for video affective content analysis. Only a subset of 40 video clips from the DEAP dataset form the basis of the experiments in [3]. Therefore, a comparison is biased towards our approach due to the increased dataset (i.e., 74 clips). On the other hand, the 40 music video clips used in [3] were selected so that only the music video clips which induce strong emotions are used. Therefore, the dataset we used in experiments is more challenging. Another difference with the setup of the work [3] is that they used the user ratings from laboratory experiments instead of the online self-assessment ratings mentioned in Section IV-A as the ground-truth. We can conclude that trained classifiers are able to better discriminate between videos with varying affective content by using MLR audio, motion, and static visual representations. Concerning the effect of dense trajectory based motion information, including the motion features further improved the performance of the method.

TABLE II. CLASSIFICATION ACCURACIES ON THE DEAP DATASET (WITH AUDIO-VISUAL REPRESENTATIONS)

Method	Accuracy (%)
Our method (MLR audio, motion & static visual) - SVM fusion	**66.22**
Our method (MLR audio, motion & static) visual) - Linear fusion	**58.11**
Our method (MLR audio & static visual) [9]	50.00
The LLR audio-visual & MLR motion	48.65
The LLR audio-visual	39.19
Yazdani et al. [3]	36.00

In Figure 2, the confusion matrices of the classification results of our method (with and without motion decision-level fusion) for the DEAP dataset are illustrated. The confusion



(a) (b)

Fig. 2. Confusion matrices on the DEAP dataset with MLR audio, motion and static visual representations (Mean accuracy: 50% for MLR audio and static visual [9] and 58.11% for linearly fused MLR audio, motion and static visual). (*ha-hv*: high arousal-high valence, *la-hv*: low arousal-high valence, *la-lv*: low arousal-low valence, *ha-lv*: high arousal-low valence). Lighter areas along the main diagonal correspond to better discrimination.

matrix on the left (Figure 2(a)) represents the performance of our method with CNN-generated audio-visual representations fused linearly at the decision-level, while the confusion matrix on the right (Figure 2(b)) represents the performance of our method with CNN-generated audio-visual and mid-level motion representations also fused linearly at the decision-level. These show the results for linear fusion only, as we did not perform SVM-based fusion in our previous work [9]. The detailed definition of the labels presented in Figure 2 is given in Section IV-A. One final observation is that including motion information improves particularly well the classification results for the quadrant of *HA-HV* and *LA-LV*. However, it does not help improving results for the quadrant of *HA-LV*. The most difficult affect quadrant to discriminate is the quadrant of *LA-HV*, whereas the method performs well for the other three quadrants. Overall results suggest that incorporating high-level representations such as sentiment-level semantics is necessary to further improve the classification performance.

V. Conclusions

In this paper, we presented an approach for the affective labeling of music video clips, where higher level representations were learned from raw data using CNNs and fused with dense trajectory based motion features at the decision-level. MFCC was employed as audio features, while color values in the HSV space formed the static visual features as a basis for feature learning. We utilized the mid-level audio-visual representations to classify each music video clip into one of the four quadrants of the VA-space using multi-class SVMs. Experimental results on a subset of the DEAP dataset support our assumptions (1) that higher level audio-visual representations learned using CNNs are more discriminative than low-level audio-visual representations and (2) that including dense trajectories contribute to increase the classification performance. As future work, we plan to concentrate on the modeling and representation aspects and explore machine learning techniques such as ensemble learning to obtain better classification performance. In addition, we aim to extend our approach to user-generated videos. Different from music video clips, these videos are not professionally edited, e.g., in order to enhance dramatic scenes. Thus, focusing on such content will shed light on the significance of actual sounds and visuals that are produced in real-world scenes.

Acknowledgments. The research leading to these results has received funding from the European Community FP7 under grant agreement number 261743 (NoE VideoSense).

References

[1] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[2] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Trans. on Multimedia*, vol. 12, no. 6, pp. 523 –535, oct. 2010.

[3] A. Yazdani, K. Kappeler, and T. Ebrahimi, "Affective content analysis of music video clips," in *MIRUM*. ACM, 2011, pp. 7–12.

[4] M. Xu, J. Wang, X. He, J. Jin, S. Luo, and H. Lu, "A three-level framework for affective content analysis and its case studies," *MTAP*, 2012.

[5] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *PAMI*, vol. 35, no. 8, pp. 1798–1828, 2013.

[6] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *PAMI*, pp. 221–231, 2013.

[7] T. Li, A. Chan, and A. Chun, "Automatic musical pattern feature extraction using convolutional neural network," in *Int. Conf. Data Mining and App.*, 2010.

[8] E. Schmidt, J. Scott, and Y. Kim, "Feature learning in dynamic environments: Modeling the acoustic structure of musical emotion," in *ISMIR*, 2012, pp. 325–330.

[9] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *MMM*, 2014, pp. 303–314.

[10] H. Wang and L. Cheong, "Affective understanding in film," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp. 689–704, 2006.

[11] P. Valdez and A. Mehrabian, "Effects of color on emotions," *Journal of Experimental Psychology: General*, vol. 123, no. 4, p. 394, 1994.

[12] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[13] J. Eggink and D. Bland, "A large scale experiment for mood-based classification of tv programmes," in *ICME*, 2012, pp. 140–145.

[14] M. Wimmer, B. Schuller, D. Arsic, G. Rigoll, and B. Radig, "Low-level fusion of audio and video feature for multi-modal emotion recognition," in *VISAPP*, vol. 2, 2008, pp. 145–151.

[15] Y. Baveye, J. Bettinelli, E. Dellandréa, L. Chen, and C. Chamaret, "A large video data base for computational models of induced emotion," in *ACII*. IEEE, 2013.

[16] G. Irie, K. Hidaka, T. Satou, T. Yamasaki, and K. Aizawa, "Affective video segment retrieval for consumer generated videos based on correlation between emotions and emotional audio events," in *ICME*. IEEE, 2009, pp. 522–525.

[17] L. Canini, S. Benini, and R. Leonardi, "Affective recommendation of movies based on selected connotative features," *Circuits and Systems for Video Technology, IEEE Trans. on*, vol. 23, no. 4, pp. 636–647, 2013.

[18] Y. Jiang, B. Xu, and X. Xue, "Predicting emotions in user-generated videos," in *AAAI*, 2014.

[19] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Action Recognition by Dense Trajectories," in *CVPR*, Jun. 2011, pp. 3169–3176.

[20] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[21] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[22] T. Wu, C. Lin, and R. Weng, "Probability estimates for multi-class classification by pairwise coupling," *The Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.