HAL
open science

Depression Estimation Using Audiovisual Features and Fisher Vector Encoding

Varun Jain, James L. Crowley, Anind Dey, Augustin Lux

To cite this version:

Varun Jain, James L. Crowley, Anind Dey, Augustin Lux. Depression Estimation Using Audiovisual Features and Fisher Vector Encoding. ACM Multimedia 2014, Nov 2014, Orlando, FL, United States. 10.1145/2661806.2661817 . hal-01081358

HAL Id: hal-01081358

https://inria.hal.science/hal-01081358

Submitted on 7 Nov 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Depression Estimation Using Audiovisual Features and Fisher Vector Encoding

Varun Jain
INRIA
Univ. Grenoble Alpes, LIG,
F-38000 Grenoble, France
CNRS, LIG, F-38000
Grenoble, France
varun.jain@inria.fr

James L. Crowley
Univ. Grenoble Alpes, LIG,
F-38000 Grenoble, France
CNRS, LIG, F-38000
Grenoble, France
INRIA
james.crowley@inria.fr

Anind K. Dey
Human-Computer Interaction
Institute, Carnegie Mellon
University
Pittsburgh, PA, USA
anind@cs.cmu.edu

Augustin Lux
Univ. Grenoble Alpes, LIG,
F-38000 Grenoble, France
CNRS, LIG, F-38000
Grenoble, France
INRIA
augustin.lux@inria.fr

ABSTRACT

We investigate the use of two visual descriptors: Local Binary Patterns-Three Orthogonal Planes(LBP-TOP) and Dense Trajectories for depression assessment on the AVEC 2014 challenge dataset. We encode the visual information generated by the two descriptors using Fisher Vector encoding which has been shown to be one of the best performing methods to encode visual data for image classification. We also incorporate audio features in the final system to introduce multiple input modalities. The results produced using Linear Support Vector regression outperform the baseline method[16].

Categories and Subject Descriptors

I.4.9 [**Image Processing and Computer Vision**]: Applications; I.5.4 [**Pattern Recognition**]: Applications—*Computer Vision*

Keywords

Local Binary Patterns; Dense Trajectories; Multimodal Affect Sensing

1. INTRODUCTION

Depression is a serious mental disorder involving persistent bad mood, low self-satisfaction and lack of interest in normal pleasurable activities. Currently depression is diagnosed by a patient's self report or through a mental status examination (MSE). A MSE entails the observation of a patient's state of mind by a psychologist to assess aspects such as attitude, mood, affect and speech. An automated system to detect depression can help both the doctors and patients with diagnosis and treatment monitoring. Such a system will also help to overcome the problem of subjective bias associated with self-reports and MSE.

Facial expressions, eye gaze and head motion are important visual features used by psychologists to gauge depression in patients. Advances in the field of computer vision allow us to automatically observe these visual features. However most research has focused on static images and posed facial expressions. Techniques that work well for posed emotions may not work well for spontaneous expressions [15]. In [2], the authors underline the importance of spatio-temporal information for affect sensing. In the work presented here, we extract spatio-temporal information using two different visual descriptors from videos of people with depression and use linear support vector machines to quantify the level of depression.

Section 2 provides a brief description of the Local Binary Patterns using Three Orthogonal Planes (LBP-TOP) [19]. LBP-TOP is a spatio-temporal extension to the popular Local Binary Patterns (LBP). As with LBP, LBP-TOP is computationally simple to compute, yet efficient at describing texture. In addition, LBP-TOP provides relative invariance to illumination changes. In this work, we have used LBP-TOP features to describe the dynamic texture of the facial region.

Section 3 describes dense trajectories. Computing dense trajectories involves 3 steps: (1) sampling feature points in a dense grid in each video frame. (2) tracking feature points using optical flow. (3) extracting features aligned with the trajectories to characterize shape, appearance and motion. It has been shown in [18] that dense trajectories outperform the state-of-the-art spatio-temporal interest points (STIP) [9] at action recognition. We use dense trajectories to capture the visual information associated with

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MM '14 November 07 2014, Orlando, FL, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3119-7/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661806.2661817.

macro-movements such as those of the head, shoulders and other parts of the upper-body visible in the videos.

In section 4 we discuss Fisher Vector encoding (FV) [13]. FV encapsulates first and second order differences between the pooled local features and the dictionary which is built using Gaussian Mixture Models (GMM). The reason why we need to encode the features and construct a signature to characterize the videos is that the video clips are of different duration. It also makes it easier for us to combine the information extracted from the two different descriptors, LBP-TOP and dense trajectories, since there are instances in the video when the face is not detected and therefore no LBP-TOP features are calculated but dense trajectories are still computed.

Our results are presented in section 5. This section discusses how the videos are pre-processed and faces are aligned. It describes the experimental protocol adopted and how the parameters are optimized. This section presents a prediction error that is lower than that of the baseline method and concludes with discussion of some of the insights gained during the experimentation.

2. RELATED WORK

In [6], the authors looked at the change over time in severity of depression and facial expressions. Facial expressions were analyzed using FACS and it was found that when the patient was suffering from severe depression, the facial expressions were consistent with the "social risk hypothesis" which states that patients with depression tend to withdraw from society. This work validated the use of automated facial expression analysis for behavioral science.

Scherer *et al.* [14] recognize vertical head gaze, vertical eye gaze, smile intensity and smile duration as important features for sensing psychological disorders such as anxiety and depression. They employ a multimodal sensor framework called Multisense which included a face tracker, a head tracker, a system for observing eye gaze and a Microsoft kinect sensor for skeleton tracking and audio capture. They discovered that people with depression generally have a downward angle of gaze as compared to non-depressed people. It was also found that depressed people have lower intensity smiles and have shorter duration smiles, on average. These findings suggest that head pose and facial expressions are important visual cues for depression.

Without using subject specific Active Appearance Models (AAM) as in [4] and [11], Joshi *et al.* in [8] use LBP-TOP and STIP features in conjunction with Bag of words (BoW) encoding to detect depression. They experiment with a variety of feature fusion techniques and combine audio features such as loudness, pitch, intensity and Mel-frequency cepstral coefficients (MFCC) to develop a multimodal depression sensing system.

3. LBP-TOP FEATURES

In [19] Zhao and Pietikäinen present a spatio-temporal extension to LBP. They propose concatenating local binary patterns on three orthogonal planes: XY, XT and YT where XT and YT contain the space-time transition information. Using uniform patterns the length of the feature vector for each plane is limited to just 59 values, leaving us with a 3 X 59 histogram for a video sequence. Unlike the circular sampling in conventional LBP, LBP-TOP uses elliptical sampling to fit to space-time statistics. Fig.1 illustrates how LBP-TOP features are calculated.

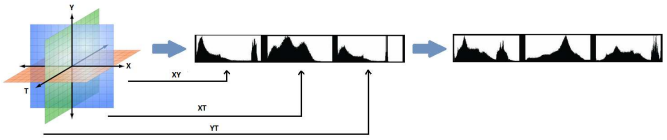

Figure 1: LBP-TOP computation

LBP-TOP features are a computationally efficient yet simple approach to describe dynamic facial texture. We use these features to capture intra-face movements.

4. DENSE TRAJECTORIES

Space Time Interest Points (STIP) introduced by Laptev [9] by extending the Harris detector to the space-time domain is a very successful approach for activity recognition. Recently Wang *et al.* in [18] demonstrated that activity recognition performance can be increased by treating the space and time domains separately, rather than detecting interest points in a joint 3D space. In contrast to STIP, dense trajectory computation involves tracking densely sampled points from each frame using an optical flow algorithm and thereby capturing motion information of trajectories.

Dense sampling of feature points carried out on multiple spatial scales provides coverage of all spatial positions. Feature points are tracked on each spatial scale separately using optical flow fields. Apart from the trajectory shape information, Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH) descriptors are computed over 3D space-time volumes aligned with the trajectory to provide motion information. Fig.2 shows how dense trajectories are computed.

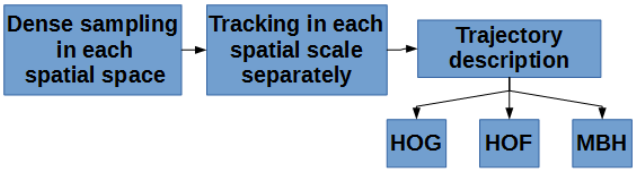

Figure 2: Dense Trajectory computation

Dense Trajectories are a state-of-the-art technique for video description based on optical flow fields. We make use of dense trajectories to extract visual information related to macro level movements.

5. FISHER VECTOR ENCODING

The Bag of Visual Words (BoV) is a vector of occurrence counts of a vocabulary of local image features constructed using off-line k-means clustering on a large set of local descriptors. Bag of word representation counts the number of descriptors assigned to a particular cluster, Fisher vector encoding not only provides that information but also encodes the deviation of a sample from the distribution in the form of first and second order statistics. Although the encoding

generated using Fisher Vectors is not sparse, the number of clusters required in Fisher Vectors for attaining an accuracy similar or better than sparse coding and bag of words is about 1/10th the number of clusters needed in sparse coding and bag of words.

Fisher Vector encoding (FV) is a major improvement over the Bag of Words (BoW) technique and sparse coding[10]. Chatfield *et al.* in [3] report that Fisher Vector encoding works better than a variety of encoding techniques at image classification on the PASCAL VOC challenge[5].

Unlike BoW where k-means clustering is used to build the dictionary, FV uses GMM for building the dictionary. This dictionary built using GMM can be visualized as a probabilistic visual dictionary. Using this GMM-based dictionary, weighted measures of the descriptor are assigned to multiple clusters in contrast to BoW encoding where descriptors are assigned to a single cluster. In this paper we use the VLfeat implementation [17] of Fisher Vector encoding which is available for free download from www.vlfeat.org.

6. EXPERIMENTS AND RESULTS

The AVEC 2014 challenge dataset is divided into 3 partitions: training, development and test. The labels for the training and development set are available to participants. Results are mailed to the organisers in order to obtain the errors over the test set. Participants get 5 attempts to test their results on the test set. Each partition contains 50 Beck Depression Index-II labels. Each label corresponds to a pair of videos, Freeform and Northwind.

To extract the visual information from the videos we compute LBP-TOP features to capture the intra-face movement and dense trajectories to capture macro movements such as those of the head and the shoulders.

We split the videos into individual frames and perform face detection and alignment using Openimaj[7]. Openimaj normalizes the detected face into an imagette of 80 X 80 pixels. Zhao and Pietikäinen [19] demonstrate that it is best to divide the imagette into overlapping spatial regions and calculate the LBP-TOP features separately for each spatial region over a time slice and finally concatenate the results from the different regions and time slices. This technique helps encode the occurrence of micro-patterns and their relative locations in the image.

Using cross-validation, it is seen that for an imagette of our size, spatial regions of 10 X 10 pixels work best with a 50% overlap. We compute LBP-TOP features for 2 different sizes of the temporal slices, t=3s and t=1s.

Principal Component Analysis (PCA) is performed to reduce the dimensionality of the LBP-TOP feature vector and decorrelate the features. This reduces the computation time and also reduces the size of the Fisher Vectors as this is linearly dependent on the feature vector size [12]. We choose a dimensionality of D=64, assuring that the variance in the projected data is at least 95% of the original data. A Gaussian mixture model is fit over a subset of reduced-dimensionality training features which is used to create one Fisher vector per video. The optimum number of clusters is chosen using cross-validation over the development set.

For dense trajectories we use the following settings: length of the trajectory is set to 15 frames, the stride for dense sampling feature points is set to 5 pixels and the neighborhood size for computing the descriptor is set to 32 pixels. A set of features (HOG, HOF and MBH) over the dense trajectories is generated for each video. Just as with the LBP-TOP features, PCA is performed followed by fitting of a GMM over a subset of projected trajectory features from the training set. Millions of trajectories are generated for the training set alone, a subset of $3.6X10^5$ is used to fit the GMM. The fitted model is used to generate a Fisher vector for each video.

The low level descriptors (LLD) audio features provided with the AVEC 2014 database are reduced in dimensionality using PCA (D=64) and a GMM is fit over the projected features followed by Fisher vector generation.

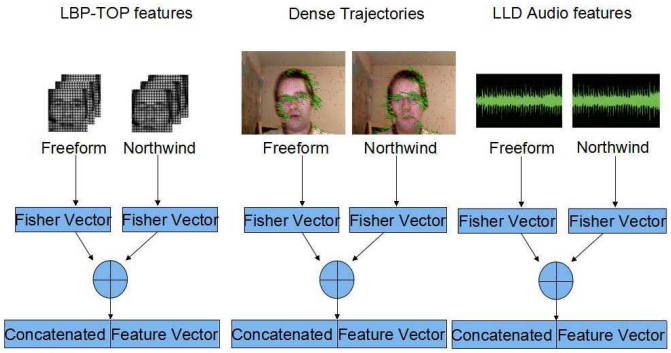

Figure 3: System Architecture

The LBP-TOP, dense trajectory and audio features having been transformed into Fisher vectors are concatenated for each pair of videos, Freeform and Northwind, and fed as input to a linear support vector machine (SVM). Linear SVM is chosen because in the feature matrix, the number of columns, are much more than the number of rows. A feature vector of D=64 produces a FV of 4096 columns, for each pair of videos there will therefore be 2*4096=8192 columns whereas the number of samples in the training set is just 50. Fig.4 shows how the optimum number of clusters is chosen for fitting the GMM on LBP-TOP features. The minimum development error is achieved at 35 clusters; a similar analysis gave us a minima at 40 for dense trajectories and at 50 for LLD audio features.

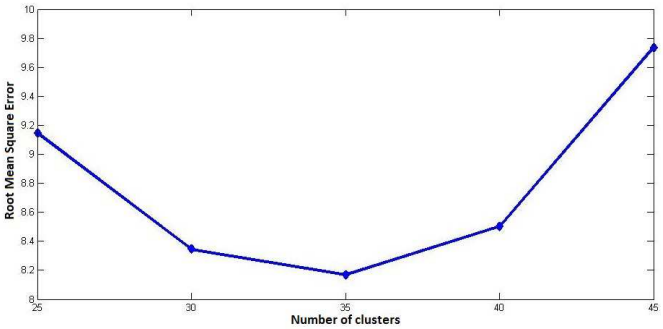

Figure 4: Root Mean Square Error (RMSE) vs. number of clusters

We compare our results obtained using Fisher vector encoding with results produced using sparse coding in Table 2 for LBP-TOP features on the development set.

	LBP-TOP	LLD	Dense Trajectories	LBP-TOP+Dense Trajectories	LBP-TOP+LLD	LLD+Dense Trajectories	LBP-TOP+LLD+Dense Trajectories
MAE	6.9697	9.7457	9.8668	6.9679	6.9662	9.5229	**6.9643**
RMSE	8.1674	11.514	11.7985	8.1647	8.1645	11.2538	**8.1618**

Table 1: Errors for different combinations of descriptors on the development set

Encoding Technique	Fisher Vector Encoding	Sparse Coding
MAE	6.9697	10.1785
RMSE	8.1674	11.9858

Table 2: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for different encoding techniques

For sparse coding, we vary the dictionary size from 250-750 and the minima is attained at 550. We use max pooling in the final encoding step which has been shown to perform better than average pooling.

Window Size	1 second	3 seconds
MAE	7.5520	6.9697
RMSE	8.9025	8.1674

Table 3: Errors for different sizes of time slice

It can be seen in Table 3 that a time slice of 3 seconds works better than a slice of 1 second on the development set.

In Table 1 we see that the minimum errors are obtained by combining all the three descriptors: LBP-TOP, dense trajectories and LLD audio features. However using LBP-TOP features alone and encoding them using Fisher Vectors, we achieve error values very close to the error values attained by the combination of all three features. Given the computational effort required to generate and encode dense trajectories and LLD features, we opt to use LBP-TOP features alone for our results on the test set.

We only test the early fusion technique because we only have 50 samples for training and another 50 for development. In case late fusion is performed, we will need two layers of regressors with the output of one layer forming the input for the second and the training data of just fifty samples getting split between the two layers.

Finally we compare our errors on the development and testing set with the baseline in Table 4.

	Development Set		Test Set	
	Our Method	Baseline	Our Method	Baseline
MAE	6.9697	-	**8.3988**	8.857
RMSE	**8.1674**	9.26	**10.2491**	10.859

Table 4: Comparison of errors with baseline

Our method performs better than the baseline method. It is worth noting that these results are produced using LBP-TOP features alone combined with Fisher Vector encoding. The baseline [16] uses LGBP-TOP features [1] which are LBP-TOP features calculated over several orders of Gabor images; LBP-TOP features are therefore computationally simpler to compute. In the baseline draft paper it is not mentioned how they encode the visual information to obtain a unique feature vector corresponding to each label hence we cannot compare the computational efficiency of our system with the baseline.

7. CONCLUSION

This paper presents a multimodal system for automated depression evaluation. It presents a method to quantitatively estimate the likelihood of depression using visual features.

Our experiments show that dense trajectories and LLD features do not significantly reduce the mean absolute and root mean square errors when the features are combined with LBP-TOP features. It is seen that LBP-TOP features alone combined with Fisher Vector encoding are enough to beat the baseline.

We believe that this novel framework for depression assessment can be easily extended for predicting other slowly changing labels such as mood.

8. ACKNOWLEDGEMENTS

The work presented in this paper was partially funded by the Région Rhône-Alpes of France through the CMIRA program.

9. REFERENCES

[1] T. R. Almaev and M. F. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *ACII*, pages 356–361, 2013.

[2] Z. Ambadar, J. Schooler, and J. F. Cohn. Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science*, 16(5):403–410, 2005.

[3] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, pages 1–12, 2011.

[4] J. F. Cohn, T. S. Kruez, I. A. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. la Torre. Detecting depression from facial actions and vocal prosody. In *ACII*, pages 1–7, 2009.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[6] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, and D. P. Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *FG*, pages 1–8, 2013.

[7] J. Hare, S. Samangooei, and D. Dupplaw. Openimaj and imageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In *ACM*

Multimedia 2011, pages 691–694. ACM, November 2011. Event Dates: 28/11/2011 until 1/12/2011.

[8] J. Joshi, R. Goecke, A. Dhall, S. Alghowinem, M. Wagner, M. Breakspear, J. Epps, and G. Parker. Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on MultiModal User Interfaces*, 7(3):217–228, 2013.

[9] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[10] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, page 87, 2009.

[11] G. McIntyre, R. Göcke, M. Hyett, M. Green, and M. Breakspear. An approach for automatically measuring facial activity in depressed subjects. In *ACII*, pages 1–8, 2009.

[12] D. Oneata, J. J. Verbeek, and C. Schmid. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, pages 1817–1824, 2013.

[13] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[14] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. A. Rizzo, and L.-P. Morency. Automatic behavior descriptors for psychological disorder analysis. In *FG*, pages 1–8, 2013.

[15] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang. Authentic facial expression analysis. *Image Vision Comput.*, pages 1856–1863, 2007.

[16] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. AVEC 2014 - 3D dimensional affect and depression recognition challenge. In *4th ACM international workshop on Audio/visual emotion challenge*, 2014.

[17] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1469–1472, New York, NY, USA, 2010. ACM.

[18] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.

[19] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):915–928, 2007.