ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Bridging face and sound modalities through Domain Adaptation Metric Learning

Christos Athanasiadis, Enrique Hortal and Stylianos Asteriadis *

Department of Data Science and Knowledge Engineering
Maastricht University - the Netherlands

Abstract. Robust emotion recognition systems require extensive training by employing huge number of training samples with purpose of generating sophisticated models. Furthermore, research is mostly focused on facial expression recognition due, mainly to, the wide availability of related datasets. However, the existence of rich and publicly available datasets is not the case for other modalities like sound and so forth. In this work, a heterogeneous domain adaptation framework is introduced for bridging two inherently different domains (namely face and audio). The purpose is to perform affect recognition on the modality where only a small amount of data is available, leveraging large amounts of data from another modality.

1 Introduction

Recent advancements in the domains of machine learning and affective computing have led to the generation of powerful systems in affect analysis, mainly through the adoption of facial expression recognition [1]. Furthermore, during the last decade, the gained popularity of deep learning architectures has led to the generation of robust computer vision systems, not only for emotion recognition, but also for several applications, such as object recognition, face detection and face recognition. These advances in machine learning derived by leveraging huge datasets with which researchers are able to train and test sophisticated and highly efficient emotion classifiers [2].

The majority of techniques employ the concepts of training and testing on datasets, which consist of samples from the same feature space and with similar distributions [3]. The training dataset is utilized to obtain the classification model while the test dataset is applied for evaluation purposes. It has been shown that, when there is a significant difference between these two sets, the performance of the system decreases [4]. The differences between training and test datasets can be attributed to the dimensionality space or the distribution of the features [5]. However, the availability of enough datasets from the same domain and distribution is not always guaranteed. For example, in the case of emotion recognition, there is a plethora of available datasets derived from facial expressions that can be easily used for creating powerful emotion recognition models [6]. However, the availability of datasets from other emotion recognition modalities, such as audio or brain signals is not as rich as in the case of face

*This work was totally supported by the Horizon 2020 funded project MaTHiSiS (Managing Affective-learning THrough Intelligent atoms and Smart InteractionS) nr. 687772 (http://www.mathisis-project.eu/).

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

modality. Therefore, generating training models for emotion recognition through these modalities can be a rather challenging task and requires the generation of robust datasets. Meanwhile, the engineering of such big and complex corpora is not always a straightforward and feasible task. In order to accommodate this task, domain adaptation (or transfer learning) algorithms are fostered in many research works, in order to perform efficient classification tasks by exploiting data from modalities with rich available datasets [5]. Transfer learning eliminates the source and target domain distribution differences by performing transformation mapping of both data into a shared latent feature space. Therefore, by mitigating the source and target domain distribution inherent differences, we can transfer knowledge across modalities with different distributions and project both on a new latent domain [7].

In this work, the proposed Heterogeneous Domain Adaptation (HDA) approach is investigated with the purpose of developing an affect-augmented system through audio recognition. Having a system performing emotion recognition through audio modality but a sparse availability in audio features, the task is to make use of available datasets from face modality with the purpose of enhancing the classification performance and in order to investigate the efficiency of knowledge transfer between these domains. Having established a transformation between the two modalities, the performance of the introduced framework is tested using an SVM classifier for the target task. The main contributions of the paper are two-fold: Firstly, this paper introduces domain adaptation across modalities of different nature with a purpose of emotion recognition and, thus, proposes an association mapping for classification purposes which is applied to both cues. Secondly, Distance Metric Learning techniques are proposed in order to obtain similar distributions across different modalities, with the ultimate goal of performing affect recognition in the wild.

2 Approach

In this section, the proposed affect-augmented framework which performs audio recognition by leveraging annotated data related to face modality is analyzed. The domain spaces of those two modalities are governed by different distributions, however, both could potentially be used for performing emotion recognition either by using facial expressions or by extracting emotions from audio features. These two modalities can not be easily bridged though, because of the inherent differences in their distributions. In order to exploit the source domain dataset for the target classification task, a transformation that will bridge the two domains needs to be established.

Firstly, for the needs of the current research, the face modality is defined as the source domain X_S while the audio modality is defined as the target domain X_T. For the face modality, 3D LBP [8] features are extracted, while for the audio, frequency-based features that capture both voice quality and prosodic characteristics of a speaker (described also in more details in [8]) are used. The proposed framework can be divided into three distinct modules: 1) the feature

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

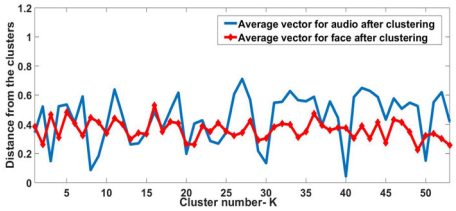

Fig. 1: Feature distributions distance after clustering.

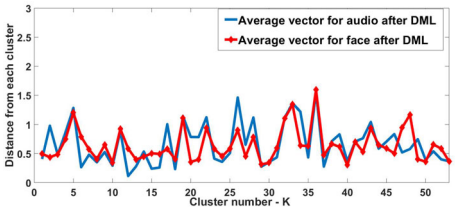

Fig. 2: Feature distributions distance after DML.

selection module, 2) the unsupervised learning module and finally 3) the Distance Metric Learning module (DML).

The first part of the approach is performed with the purpose of implementing a feature selection technique in the extracted features. This step is implemented in order to keep only the features, both from face and audio domains, which are most significant for the classification task in hand (namely, emotion recognition). For the feature selection, tree-based estimators (random forests) are used to compute feature importance, which, in turn, is used with the purpose of discarding irrelevant features. After feature selection, source and target domains are denoted as $\hat{X}_S$ and $\hat{X}_T$ respectively.

Subsequent to feature selection, an unsupervised learning technique is performed in order to code the already filtered features. During this approach, K clusters are calculated for both modalities using a clustering technique. Then, every input feature vector for both domains is transformed into the distance of each feature vector to the calculated centers ($\Phi_{K_S}(\hat{X}_S)$ and $\Phi_{K_T}(\hat{X}_T)$) through the clustering approach. That procedure is denoted as Φ_K which represents the transformation of the features to the distances from the calculated centers. The chosen approach for the calculation of the clusters that is tested in this step is the k-means.

However, even after the implementation of the clustering approach, the distributions of face and audio cannot be considered as comparable yet (as it can be seen in Fig. 1). A transformation is needed in order to perform the desired bridging between the two domains. In the current work, DML techniques are chosen in order to fill this gap.

A possible intuitive formulation for the DML problem could be as follows: given an input distance function $d(x_i, x_j)$ between objects x_i and x_j (for example, the simple Euclidean distance), together with some supervised information related to what is considered to be the ideal distance, DML's target is to formulate a new distance function $\hat{d}(x_i, x_j)$ which is more appropriate for a specific goal in comparison to the initial distance [9]. All tested methods for DML assume that we have some supervised information available. In the current study, the hypothesis that the label information from both domains can work as supervised input for calculating and optimizing the distance metric $\hat{d}$ is established. The supervised information (relation between pairs of vectors of what is considered

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

being similar or dissimilar) can be mathematically framed with the following equation:

$$C_{ij} = \begin{cases} 1 & \text{if } (x_i, x_j) \in S \\ -1 & \text{if } (x_i, x_j) \in D \end{cases} \quad (1)$$

where the sets S and D correspond to similar and dissimilar pairs of subjects. The generation of the distance $\hat{d}(x_i, x_j)$ is based on these constraints. The goal is to learn the matrix M in equation 2:

$$\hat{d}_m(x_i, x_j) \quad = \quad (x_i - x_j)^T M (x_i - x_j) \quad (2)$$

where $M = AA^T$ and it has to be a $n \times n$ positive defined and symmetric matrix [10], n is the number of input features (both for face and audio domains after the clustering coding in our case) and x_i, x_j are two feature vectors where i represent source and j target domain respectively. The matrix A^T can be used to transform the input to the new DML domain. The next step of DML is to minimize the distances between those adjacent examples indicated in Equation 1. This minimization can be framed from the following loss function:

$$L(S, D) = \frac{1}{2} \sum_{i,j}^{n} ||(x_i - x_j)^T M (x_i - x_j)|| C_{ij} \quad (3)$$

The calculated matrix A^T is used to project data in the DML space where the new distance between the pairs that are considered similar (pairs with same labels) will decrease and it will increase for the pairs that are considered dissimilar (pairs with different labels). In the current approach, the Mahalanobis distance is calculated using an equal number of pairs (of transformed features from source $\Phi_{K_S}(\hat{X}_S)$ and target domain $\Phi_{K_T}(\hat{X}_T)$) that have the same or different labels. The pairs $(x_i, x_j) \in S$ correspond to samples from audio and face modalities from the same class while the pairs $(x_i, x_j) \in D$ correspond to samples from both domains and from different classes. The projected source domain is noted as $H_S = A^T \cdot (\Phi_{K_S}(\hat{X}_S))$ while the projected target domain is noted as $H_T = A^T \cdot (\Phi_{K_T}(\hat{X}_T))$. In order to calculate A^T and generate a robust bridge between the domains of face and audio, Sparse Determinant Metric Learning (SDML) [10] was utilized. Finally, a multiclass SVM is implemented, in order to measure the classification performance in the target domain (audio) for the task of emotion recognition. This SVM classifier is trained using the transformed merged dataset (which incorporates both modalities) in the DML space.

3 Experimental results

The proposed domain adaptation framework was validated on the challenging database "Acted Facial Expressions In The Wild" (AFEW) [11]. AFEW is a popular dataset for emotion recognition where the data gathering procedure took

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Case/Sparsity	**20**%	**30**%	**50**%	**Dense**
Baseline	21.29%	24.25%	28.03%	28.84%
HDA algorithm	25.60%	26.95%	28.30%	29.11%

Table 1: Classification performance for different sparsity cases for the baseline approach and the proposed algorithm

place in uncontrolled environments. The task of the proposed framework is to perform emotion recognition through the target modality by leveraging the data of the source modality.

For validation, two different scenarios are tested. Firstly, in order to perform a broader evaluation scheme, a test for the calculation of the projection matrix A^T was performed by incorporating the full source X_S and target domain X_T (and performing the feature selection and clustering steps as well). The number of samples from both modalities needed to be identical when calculating A^T. The reason behind this is the fact that DML is fed with pairs of instances from both modalities. Secondly, we tested a scenario consisting of sparse data from the target domain and dense information from the source domain. In this case, instead of using the complete version X_S and X_T utilized in the dense scenario, we conducted experiments considering sparse pairs from both modalities (X_{S1}, X_{T1}) in order to learn A^T. For feature selection and clustering, X_{T1} and X_S sets were used. Subsequently, following a 4-fold cross validation, pairs of samples (X_{S1}, X_{T1}) that correspond to certain percentages of the whole dataset (namely, 20%, 30% and 50%) were randomly selected with the scope to learn A^T. That matrix was used to map the training dataset from the source modality X_S and the sparse training dataset from the target modality X_{T1} into the common DML space. Then, both datasets were merged and a new training dataset emerges, $H_{train}^{Sparse} = (H_S, H_{T1})$ (while $H_{train}^{Dense} = (H_S, H_T)$ respectively).

An exhaustive search was performed to find the ideal number of clusters, for the DML parameters [10] and SVM parameters for both scenarios. Furthermore, SVM classifiers were trained on cluster-based transformed audio features (it was considered as baseline method) without taking advantage of the source domain, in order to compare them with the proposed technique. For the sparse scenario, the training set was $B_{train}^{Sparse} = \Phi_{K_{T1}}(\hat{X}_{T1})$ while for the dense scenario was $B_{train}^{Dense} = \Phi_{K_T}(\hat{X}_T)$. The prediction was performed in the transformed available test dataset (solely from the audio modality $H_{T_{test}}$ which is the same for both scenarios).

As it is depicted in Table 1 and Fig. 2, the whole framework succeeded in the objective of improving the classification performance in the target domain by incorporating information from the source dataset (and outperformed the baseline method) in all scenarios. A striking observation extracted throughout the experimental phase was the fact that the proposed method (for each sparsity case) always converged to the same classification performance using the same number of clusters and sparsity parameters. For the baseline approach, due to the randomness (in picking samples for the sparse scenario) we needed to

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

iteratively perform multiple classifications in order to reach a safe estimation of the result. This observation denotes the efficiency for the knowledge transfer from face to audio during the proposed domain adaptation approach.

4 Conclusion

In this paper, a domain adaptation framework is implemented for the challenging task of emotion recognition through audio, by incorporating data from face modality. The study focuses on the capability of domain adaptation when using two different modalities for emotion recognition which derive from different feature spaces and distributions. The goal is to analyze the linkage between modalities and mitigate the gap between their inherent distribution differences. The proposed approach incorporated an unsupervised learning step and a metric learning technique with the purpose of establishing a bridge between both domains. The performance of the proposed HDA algorithm outperformed that obtained using a baseline SVM algorithm trained solely with the target domain features. Therefore, in the experimental phase, it can be concluded that the proposed approach can successfully bridge the two inherently different domains.

References

[1] C.A.Corneanu, M.Oliu, J.F.Cohn and S.Escalera, Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect related Applications, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.

[2] Y.LeCun, Y.Bengio and G.Hinton, Deep learning, Nature Vol 521, P. 436-444, 2015.

[3] S.B.Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Procs. of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in e-Health, HCI, Information Retrieval and Pervasive Technologies, P. 3-24, 2007.

[4] H.Shimodaira, Improving predictive inference under covariate shift by weighting the log likelihood function, of Statistical Planning and Inference, Vol. 90, Iss. 2, 2000.

[5] K.Weiss, T.Khoshgoftaar and D.Wang, A survey of transfer learning, Transactions of Big Data, 2016.

[6] R.Sun and E.Moore, A Preliminary Study on Cross-Databases Emotion Recognition using the Glottal Features in Speech, Procs. ISCA's 13th, Portland, 2012.

[7] S.J.Pan and Q.Yang, A Survey on Transfer Learning, IEEE Transactions on Knowledge and Data Engineering, 2010.

[8] E.Ghaleb, M.Popa, E.Hortal and S.Asteriadis, Multimodal Fusion Based on Information Gain for Emotion Recognition in the Wild, IEEE Intelligent Systems Conference, London, UK, 2017.

[9] F.Wang and J.Sun, Survey on distance metric learning and dimensionality reduction in data mining, Transactions on Data Mining and Knowledge Discovery, Vol. 29, Iss. 2, 2014.

[10] Q.G.Jun, J.Tang, J.T.Zha, T.S.Chua, and H.J.Zhang, An Efficient Sparse Metric Learning in High dimensional Space via L1 penalized Log determinant Regularization, 26-th Annual International Conference on Machine Learning, ICML '09, P. 841-848, 2009.

[11] A.Dhall, R.Goecke, S.Lucey, and T.Gedeon, Collecting large, richly annotated facial expression datasets from movies, IEEE Transactions of MultiMedia, Vol. 19, Iss. 3, 2012.