Cross-Domain Personality Prediction: From Video Blogs to Small Group Meetings

Oya Aran[1] and Daniel Gatica-Perez[1,2]
[1] Idiap Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
(oaran,gatica)@idiap.ch

ABSTRACT

In this study, we investigate the use of social media content as a domain to learn personality trait impressions, particularly extraversion. Our aim is to transfer the knowledge that can be extracted from conversational videos in video blogging sites to small group settings to predict the extraversion trait with nonverbal cues. We use YouTube data containing personality impression scores of 442 people as the source domain and a small-group meeting data from a total of 102 people as our target domain. Our results show that, for the extraversion trait, by using user-created video blogs, as part of the training data, and a small amount of adaptation data from the target domain, we are able to achieve higher prediction accuracies than using only the data recorded in small group settings.

Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: Pattern Recognition Applications—*Signal Processing, Computer Vision*;
J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology, Psychology*

General Terms

Algorithms, Theory

Keywords

Domain adaptation, personality prediction, nonverbal behavior, social interaction

1. INTRODUCTION

One of the challenges of analyzing human behavior in social contexts is the collection of natural human behavior data. The data should be collected in a way that does not destroy the naturality of the behavior and at the same time should be suitable for automatic audio-visual processing. Due to these restrictions, most collected data sets are generally of moderate sizes. However, unlike the limited amount of data that is used to build computational models of face-to-face behavior, social media sites provide a vast amount of behavioral data. In this study, we are interested in using social media content to learn models of personality traits of individuals during interaction in small-group settings.

Predicting personality using automatically extracted nonverbal cues has been addressed in several recent studies. While some works investigated personality in small-group settings [8], other works looked at monologue-like presentations [1, 2]. One of the novelties of our work is the use of video blogging as a large-scale source of conversational data for learning models of personality traits to be transferred to other settings. In video blogs (vlogs), people talk to the camera as if they were talking to other people [13]. This results in the display of natural conversational behavior for a variety of situations and with a wealth of nonverbal communicative cues not available in other video data sources in terms of scale and diversity. As the small group domain (i.e. groups containing 3-6 people interacting face-to-face) is different than the domain of the social media data, we investigate ways to combine the two domains and assess the cross-domain prediction performance.

Classical machine learning techniques assume that the training data and the test data come from the same domain and from the same distribution, however this is rarely the reality in practice. While many real-world applications are susceptible to this fact, it is particularly true for the analysis of personality: the domain strongly determines the traits that could be encoded and observed [5]. To find a solution to this problem, domain adaptation techniques have attracted interest in recent years [11]. Based on the type and nature of the domains, i.e., the availability of labeled data, the extracted features, and the tasks that will be performed in each domain, the problem takes a different form and is called transfer learning, multitask learning, covariate-shift, etc. [10]. Our study presents the problem in a form where the source and target domains are different but related, the feature space is homogeneous, and the tasks are the same.

We apply and compare several domain adaptation approaches to perform cross-domain personality prediction, for predicting extraversion impressions in a classification task. We make use of YouTube videos as our source domain. To our knowledge, this is the first work to perform domain adaptation in personality prediction from perceptual cues. The multimodality aspect of this study comes from the use of multiple domains. To focus on multiple domains and to emphasize the cross-domain performance, we used only vi-

ICMI'13, December 9–12, 2013, Sydney, Australia
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522858 .

sual nonverbal features, represented by visual body activity statistics.

2. DATA AND ANNOTATIONS

We use two datasets from two different domains to study the cross-domain performance of personality impression prediction. Our main aim is to perform prediction in small group meetings, which is our target domain. As our source domain, we use a dataset of conversational vlogs downloaded from YouTube. For personality ratings, we use personality impressions from external observers in both cases.

2.1 Target domain: small group meetings

As our target domain, we used a subset from the Emergent LEAder corpus (ELEA)[12] for this study. The ELEA AV subset consists of audio-visual recordings of 27 meetings, in which the participants perform a winter survival task with no roles assigned. There are 102 participants in total (six meetings with three participants and 21 meetings with four participants). Each meeting lasts around 15 minutes and is recorded with two webcams and a microphone array. More details about the data can be found in [12].

For this study, we collected personality impressions of external observers watching the participants in the ELEA AV corpus. We used the Ten Item Personality Inventory (TIPI), with a 7-point Likert scale, for measuring the Big Five traits of the participants [6]. For each participant, we selected a one-minute segment from the meeting, which corresponds to the segment that includes the participant's longest speaking turn. This segment is selected as the participants are typically more expressive and more active when speaking, conveying more nonverbal cues to observers. We isolated the video of each participant such that only a single participant is visible. Figure 1(a) shows a snapshot from the videos used in the annotations. The videos are shown muted to cancel out any effects resulting from the meeting language. Each video is annotated by three different annotators and a total of five annotators annotated the whole dataset.

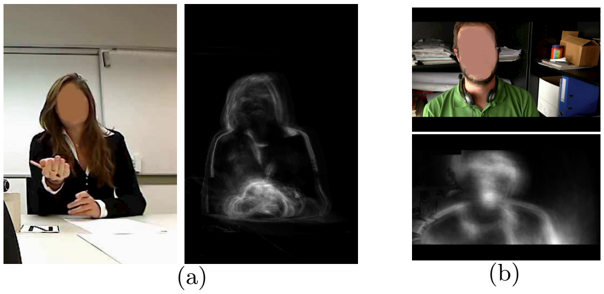


Figure 1: A snapshot from one of the annotated videos and its corresponding wMEI on (a) ELEA corpus, and (b) VLOG corpus.

2.2 Source domain: video blogs

As the source domain, we used a dataset of conversational vlogs downloaded from YouTube, first presented in [3]. We used a subset of this data, which was annotated for Big Five personality impressions, in [2]. The subset contains one video per user, resulting in a total of 442 vlogs. For the personality annotations, the first conversational minute of each vlog was obtained and shown to the annotators. The annotations were collected on Amazon Mechanical Turk and the TIPI questionnaire was used to obtain the personality impressions. Each vlog was annotated by five annotators. More details about the annotations can be found in [2]. Figure 1(b) shows a snapshot from the videos used in the annotations.

3. OUR APPROACH

3.1 Analysis of the Source and Target Domains

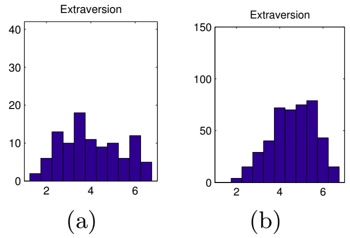
The target and the source domains that we selected to use for personality impression prediction have different properties. The source domain includes vlogs which are mainly monologues of vloggers, and as a result, the person in the video is always at the focus, having the floor. While these monologues have similarities with face-to-face conversation, the information flow is one-way. On the other hand, the target domain includes recordings of small group meetings. The participants are in an interaction and a participant may not always be at the focus. Even though for the annotations we have selected the segment in which the participant is verbally active, some participants take the floor, speak for a brief amount of time, and then, leave the floor to some other person in the rest of the segment.

For each participant, the overall personality impression score for the extraversion trait is obtained by calculating the average of the scores of the annotators. The distribution of the average extraversion scores is shown in Figures 2(a) and 2(b), for the ELEA and VLOG datasets, respectively. The plots show that the extraversion trait has a flat distribution in the ELEA corpus, while extraversion in the VLOG corpus has slightly higher mean and a peaked distributions.

We selected the extraversion trait as the focus of this study, as it is one of the strongest encoded traits during face-to-face conversations [5]. In our datasets, extraversion also receives the highest agreement between the annotators (Intra Correlation Coefficient, ICC(1,k), is 0.73 and 0.77, respectively for ELEA and VLOG datasets). The mean value of the extraversion trait is 4.06 and 4.61 in the ELEA and VLOG corpus, respectively, on a scale of 1 to 7. The difference between the means is significant (with $p = 4e^{-6}$ using a two sample t-test), which is an expected outcome given the inherent properties of the two domains. The VLOG corpus contains people who choose to record and broadcast video of themselves, which will be watched by many unknown people. It is not surprising to see that the people in this corpus are scored high in extraversion in comparison to the people in the ELEA corpus, which contains recordings of small group meetings in a laboratory setting. The participants in the ELEA corpus might represent a more "general" population and less biased towards extraverted people.

3.2 Nonverbal Feature Extraction

Extraverted people are known to be more expressive: they use a louder voice, use a larger body area by extending their arms and hands, and use more energetic and frequent gestures [9]. In this study, we have focused only on visual activity to extract nonverbal features for several reasons: First, visual body activity is one of the key nonverbal cues to signal extraversion. Second, visual activity is a more robust feature given the properties of the two domains (i.e. audio turn taking behavior in the two domains is highly different.). As the annotators have seen only a one-minute segment per



(a) (b)

Figure 2: Histogram of the Extraversion scores on (a) ELEA corpus, and (b) VLOG corpus.

participant, both for the ELEA and VLOG dataset, we have processed the same segment to extract the features.

We have used weighted Motion Energy Images (wMEI) [12] as descriptors of spatio-temporal body activity and calculated the wMEI of each video. Sample wMEIs for each corpus are shown in Figure 1. We extracted several statistics from the wMEIs, such as mean, median, 75% quantile, and entropy. As additional features, mean, median, and quantile statistics are also calculated by omitting zero intensity pixels in wMEI. For normalization, we used the maximum accumulated pixel value of the wMEI.

3.3 Domain adaptation

We investigate the use of different domains for training models for extraversion impression prediction. We are interested on the performance of predicting extraversion in small group settings, thus we use the source domain only during the training phase: the labeled training data comes from both domains while the test examples are always selected from the target domain, which is the ELEA dataset. In Figure 3, we show the flowchart of our approach, for the training and test phases separately.

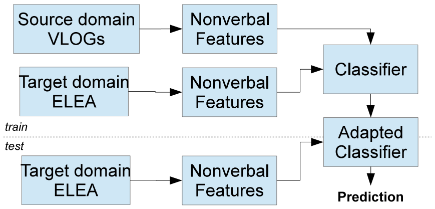


Figure 3: Flowchart of our approach for predicting extraversion impression with domain adaptation.

Below, we present the approaches that we used to build models that utilize the source and target domains for predicting extraversion. For each approach, we used two different classifiers: a ridge regression classifier and a Support Vector Machine (SVM) with a linear kernel [7].

The target only approach (**TRG**) uses only the data coming from the target domain as the training data. It is the traditional single domain approach and it provides a baseline for the domain adaptation models.

The source-only approach (**SRC**) simulates the case where there are no labeled training data from the target domain. The models are trained using the source domain data only and tested on the examples from the target domain.

The combined approach (**COMB**) uses the union of source and target data in the training, assuming that labeled training data from the target domain is available. We combine the available training data from the two domains into one training set and train the models with this combined set.

Rather than combining the features in different domains, the output of the source classifier can be used as a feature together with the target domain. We take the predictions (**PRED**) of the SRC classifier and append them to the features extracted in the target domain and train a new classifier in the augmented domain.

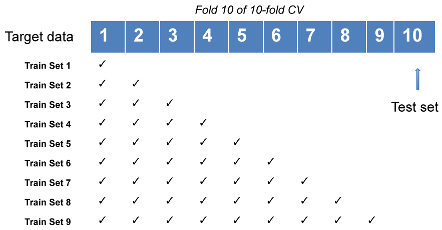
For classifiers that produce a decision score in addition to the classification decision, score fusion (**FUSE**) can be applied to combine the two domains. From the trained SRC and TRG models, we calculate the mean score of the two models and assign the class labels based on the fused score. We applied this only to the ridge regression classifier, using the estimated regression scores as the decision scores.

Finally, a feature augmentation (**AUG**) method [4] augments the feature space of the two domains for a better representation of the combined domain. The augmented feature space is formed from the two domains by using the mappings $\theta_s =< x, x, 0 >$ and $\theta_t =< x, 0, x >$, then a classifier is trained in the augmented feature space.

4. EXPERIMENTS AND RESULTS

4.1 Experimental Setup

To investigate the amount of target data needed to adapt the source domain to the target domain, we applied a 10-fold cross validation on the target data. The target data, ELEA, is divided into 10 segments and in each fold, one segment is reserved for the test data. From the remaining nine segments, nine training sets are formed with different sizes. The first training set uses one segment, the second training set uses two of the segments, up until the ninth training set that uses nine segments for the training data. One fold of the scheme is shown as an example in Figure 4. As a result, in each fold, there are nine training sets with different sizes, all of which are evaluated on the same test set. The cross validation folds are stratified to ensure the same class balance. To account for the differences in cross validation partitioning, we repeated the above procedure ten times and reported the average accuracy.



Fold 10 of 10-fold CV

Target data	1	2	3	4	5	6	7	8	9	10
Train Set 1	✓									↑
Train Set 2	✓	✓								
Train Set 3	✓	✓	✓							Test set
Train Set 4	✓	✓	✓	✓						
Train Set 5	✓	✓	✓	✓	✓					
Train Set 6	✓	✓	✓	✓	✓	✓				
Train Set 7	✓	✓	✓	✓	✓	✓	✓			
Train Set 8	✓	✓	✓	✓	✓	✓	✓	✓		
Train Set 9	✓	✓	✓	✓	✓	✓	✓	✓	✓	

Figure 4: Experimental setup: the test and training sets in fold 10 of 10-fold cross validation are shown.

We formulate the problem as a classification problem and assigned 0/1 labels to indicate low/high extraversion using the median value as a threshold. For training a ridge regression classifier, we used the 0/1 labels as the scores. For prediction, the trained model is used to make an estimate and the predicted label is set to 0 if the estimated score is less than 0.5, otherwise the predicted label is set to 1. Both source and target data are normalized separately such that each feature has zero mean and one standard deviation.

The parameters of the ridge regression and SVM with linear kernel are optimized using a nested cross-validation scheme. The ridge parameter and the C parameter of the SVM are selected from a range of $[2, 150]$ and $[2^{-5}, 2^{5}]$, respectively.

4.2 Results

The results of the experiments are shown in Figure 5. The highest mean accuracy (70.4%) is obtained by COMB with ridge regression. The accuracy changes only slightly with the amount of target data used. Even using no target domain data at the training phase, SRC achieves accuracies of 68% and 69% for ridge and SVM, respectively. The target only approach on the other hand, has the lowest accuracy among all models (with the exception of AUGM with ridge regression having the lowest accuracy up to using 40% of target domain data). The highest accuracy with TRG is 67.8%, obtained with ridge regression, by using approximately 90 examples from the target domain. The accuracies of PRED and FUSE increase with the amount of target data used, reaching to the SRC performance after using 80 and 60 target domain examples, for ridge and SVM classifiers respectively. AUGM with SVM achieves high accuracy using only a few target examples, whereas AUGM with ridge produces the lowest accuracy with the same setup.

Overall the results show that almost any attempt to use the source domain produces a higher accuracy than using only the target domain. Simple methods, such as using only the source domain (SRC) or a union of the two domains (COMB), produce the highest accuracies. PRED, FUSE, AUGM do not introduce an improvement on the accuracy in comparison to SRC and COMB. The use of larger datasets for experiments could result in observable differences.

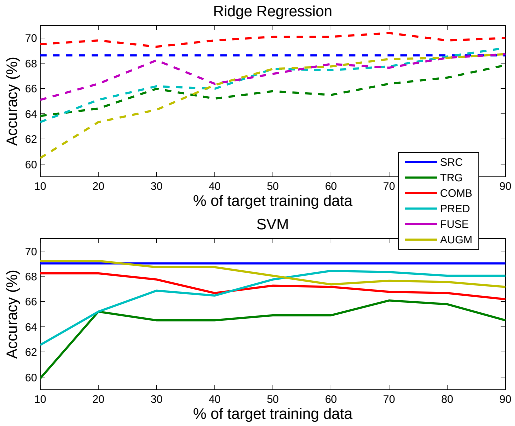


Figure 5: Accuracy of ridge and SVM classifiers with different methods with respect to the amount of labeled training target data. Best viewed in color.

5. CONCLUSIONS

Our results indicate that for the binary classification of extraversion impression, a model learned over body activity cues on vlog data can be useful in a transfer learning setting with face to face interaction in small groups as the target domain. Using a video blog data and only a small amount of data, as low as 10 examples collected from the small-group domain is sufficient for building models to predict the extraversion impressions of individuals with 70% accuracy. This shows that this data source is suitable to build models of personality impressions, which can be transferred to real-life settings, e.g. meetings. As future work, we plan to investigate other domain adaptation techniques and develop specialized techniques for the personality prediction task to increase the accuracy of the prediction. The use of other similar domains, for example interviews, and the use of other modalities for feature extraction can also be explored.

Acknowledgments

This work has been supported by the Swiss National Science Foundation (SNSF) Ambizione fellowship under the SOBE project (PZ00P2_136811) and the SNSF NCCR IM2. We thank J.I. Biel (Idiap) for sharing the YouTube data.

6. REFERENCES

[1] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. Please, tell me about yourself: automatic personality assessment using short self-presentations. In *ICMI*, pages 255–262, 2011.

[2] J.-I. Biel, O. Aran, and D. Gatica-Perez. You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In *ICWSM*, 2011.

[3] J.-I. Biel and D. Gatica-Perez. Vlogcast yourself: Nonverbal behavior and attention in social media. In *ICMI*, 2010.

[4] H. Daumé III. Frustratingly easy domain adaptation. In *ACL*, Prague, Czech Republic, 2007.

[5] R. Gifford. Personality and Nonverbal Behavior: A Complex Conundrum, *The SAGE Handbook of Nonverbal Communication*, pages 159–181. SAGE, 2006.

[6] S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37:504–528, 2003.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics., Springer, NY, USA, 2001.

[8] B. Lepri, S. Ramanathan, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe. Connecting meeting behavior with extraversion - a systematic study. *T. Affective Computing*, 3(4):443–455, 2012.

[9] R. Lippa. The nonverbal display and judgment of extraversion, masculinity, femininity, and gender diagnosticity: A lens model analysis. *Journal of Research in Personality*, 32(1):80 – 107, 1998.

[10] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, Oct. 2010.

[11] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

[12] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia*, 14(3):816–832, 2012.

[13] M. Wesch. Youtube and you: Experiences of self-awareness in the context collapse of the recording webcam. *Explorations in Media Ecology*, 2009.