Predicting Viewer Perceived Emotions in Animated GIFs

Brendan Jou
Electrical Engineering
Columbia University
New York, NY 10027

Subhabrata Bhattacharya
Imaging and Computer Vision
Siemens Research
Princeton, NJ 08540

Shih-Fu Chang
Electrical Engineering
Columbia University
New York, NY 10027

{bjou,sfchang}@ee.columbia.edu, subhabrata.bhattacharya@siemens.com

ABSTRACT

Animated GIFs are everywhere on the Web. Our work focuses on the computational prediction of emotions perceived by viewers after they are shown animated GIF images. We evaluate our results on a dataset of over 3,800 animated GIFs gathered from MIT's GIFGIF platform, each with scores for 17 discrete emotions aggregated from over 2.5M user annotations – the first computational evaluation of its kind for content-based prediction on animated GIFs to our knowledge. In addition, we advocate a conceptual paradigm in emotion prediction that shows delineating distinct types of emotion is important and is useful to be concrete about the emotion target. One of our objectives is to systematically compare different types of content features for emotion prediction, including low-level, aesthetics, semantic and face features. We also formulate a multi-task regression problem to evaluate whether viewer perceived emotion prediction can benefit from jointly learning across emotion classes compared to disjoint, independent learning.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

Keywords

affective computing; emotion prediction; animated gifs; perceived emotion; multi-task learning

1. INTRODUCTION

Our work focuses on the computational prediction, or recognition, of emotions perceived by viewers in animated Graphical Interchange Format (GIF) images. First, we explain why GIFs are an important media type for studying human emotion. Animated GIF images are a largely unexplored media in Multimedia and Computer Vision research. Their use for conveying emotions has become widely prevalent on the Web, and are now massively found on digital forums, message boards, social media, and websites of every genre. Videos, on one hand, are as popularly used for education and documentation as they are for car chases and explosions in a movie. The challenge with both general consumer videos [10] and professional video content [2, 11, 15, 14] is that this range of emotional expressiveness is too broad and a proper sampling scheme is difficult to determine. On the other hand, still images lack the temporal context necessary for observing concrete emotions. Meanwhile, animated GIFs have quickly become a channel for visually expressing emotion in our modern society. Their role in popular culture has even contributed to the rise of widely, rapidly spread cultural references called *memes*. This social function of GIFs today provides confidence that GIFs gathered from the Web for research are emotionally expressive.

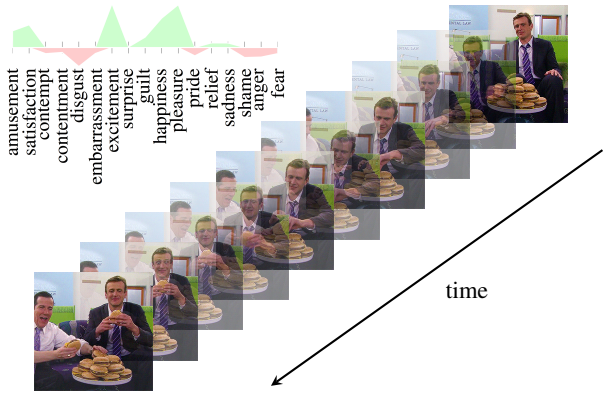


Figure 1: Example Animated GIF with Emotion Scores. Ten frames of an animated GIF sequence are shown with its respective emotion scores from user annotation on the upper-left. Scores that reflect positive presence of emotions are shown in green and negative are shown in red. The diversity of emotions illustrates the various types of emotions involved in multimedia interactions, e.g. intended, perceived, and induced.

Next, we make an explicit distinction among different types of emotions and justify the importance of perceived emotion in affective study. Affective Computing today is largely rooted in the task of recognizing human emotion (or synonymously, human affect) [13] – and more specifically, "induced" human emotion [9]. Much research in this direction follows a canonical process of presenting stimuli to human subjects, measuring their physiological signals (as well as facial, voice, and body expression), and then manually or computationally analyzing the outcome [7, 15]. Subjects are often asked to complete a survey to qualitatively or quantitatively describe their emotions to the stimuli post hoc [11, 14, 15]. A key issue with this paradigm is it assumes there is only one type of emotion to study, when, in fact, *not all affect is induced*. While traditional Affective Computing has assumed a single emotion or-

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2656408.

acle as the target, we propose that there are in fact, many different types of emotion.

In Figure 1, we show an example of how there may be different categories of emotion, or affect. The example animated GIF shows two characters, one in white, one in black, are seated in front of a large pile of hamburgers. As they dig into this pile of burgers, the one in black makes a wide-eyed expression, while the character in white makes a wide-mouthed expression. Perhaps the author of this image sequence designed it for a setting where they *intended* to make their audience feel "disgusted" at gluttony of these two characters. What we show at the top-left of Figure 1 though is that viewers can *perceive* that the GIF expresses "pleasure," "happiness," "excitement," and "satisfaction" – that is, they cognitively understand that the image sequence portrays those emotions. Ultimately though, despite what they perceive, they may actually feel (or, be *induced* with) something different – "guilt" or "shame" for having done something similar before or, for others, a sense of "amusement" because they find the GIF to be funny.

Perceived emotion is an important phenomena to study because it is more concrete and objective than induced emotion, where labels are less reliable due to their subjectivity. In addition, computationally recognizing what an author truly intended is challenging because such labels often do not exist. In this work, we aim to evaluate features and computational models for predicting perceived emotions of GIFs. Specifically, we compare features of different types, including low level features like color histograms, aesthetics, mid-level semantic features inspired by emotion modeling, and face features. Also, in view of the relatedness of multiple emotions, we apply and evaluate multi-task learning, which has been designed to perform multiple machine learning tasks jointly.

The key contributions of our work include: (1) the first work to computationally predict emotion in animated GIFs to the best of our knowledge, (2) the introduction of different types of emotion, while explicitly focusing on perceived emotion, one of the first works for this emotion type, and (3) the prediction of 17 discrete emotion perceived by viewers using multi-task learning.

2. RELATED WORK

In [4], the authors propose a bank of visual classifiers that form a mid-level representation for modeling affect in images called SentiBank. Briefly, the representation is a set of 1,200 Linear SVM outputs where the SVMs are trained using a taxonomy of "adjective-noun pairs" (ANPs). The ANPs combine a "noun" for visual detectability and an "adjective" for affective modulation of the noun, resulting in pairs like "cute dog," "beautiful sunset," "disgusting food," and "terrible accident." These ANPs were mined from Flickr, where authors uploaded images along with tag metadata to describe their content. Although not acknowledged by the authors in [4], because of this way that SentiBank is trained, the representation actually describes an uploader's or originator's intended emotion, not emotion in general. Recent work in [6] studied the metadata of images and used a text-based model to extract "publisher affect concepts" and "viewer affect concepts", but tries to divide emotion along the axis of human roles, e.g. publishers and viewers, but does not acknowledge that emotion can also be divided along how they arise in those humans, e.g. intended, perceived, and induced. We also set a more ambitious goal of predicting viewer perceived emotion using content-based methods where such metadata and comments are not available, as is often the case.

In Affective Computing for visual data, [12] develops a set of color, texture, and composition features inspired by concepts in art theory and psychology and uses Naïve Bayes classifiers for emotion recognition in still images. In [18], a global color-based feature

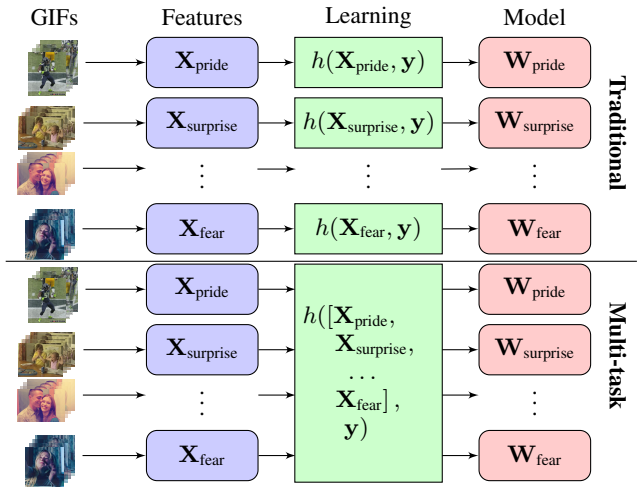


Figure 2: Traditional and Multi-task Learning. Top: Traditional pipeline for discrete emotion recognition where classifier or regressor h is learned independently for each emotion with training labels $\mathbf{y}$. Bottom: Multi-task learning jointly learns over all emotion classes yielding multiple regressors.

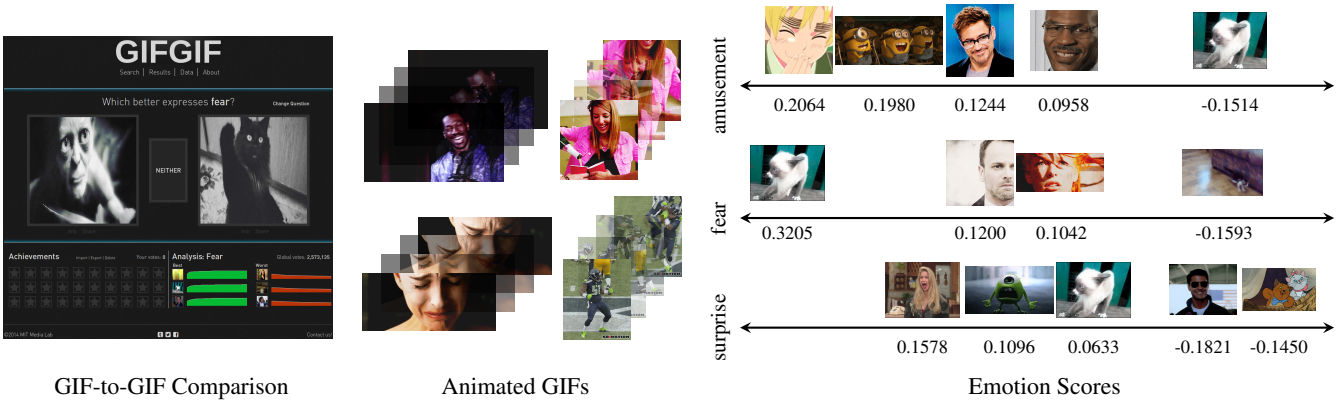
is designed for a semi-supervised factor graph formulation to predict emotion in images and modify the "feel" of an image. Both of these representations model visual emotion using early fusion and independently trained classifiers for each emotion. We use multi-task regression to learn multiple regressors jointly across emotions, resulting in a fast, low-cost linear projection at test time.

There are several existing public datasets available to the community for recognizing emotion in visual media, but all focused on induced emotion to best of our knowledge. The International Affective Picture System (IAPS) was an early dataset that does deviate slightly from induced emotion, but only provides a small set of ~1,000 still images. The DEAP dataset [11] consists of 40 one-minute excerpts from music videos and spontaneous physiological signals from 32 subjects. The MAHNOB-HCI Tagging dataset [15] has 20 short film clips with ~27 subjects with physiological signals like EEG and audio recordings of the subjects. FilmStim [14] and LIRIS-ACCEDE [2] are two other datasets consisting also of short film clips, 64 and 9,800 clips respectively, but use scores and rankings by participant ratings instead of strict classifications. These datasets use an emotion model called the "valence-arousal(-dominance) space" that represents emotion along a "valence" axis measuring sentiment positivity and an "arousal" axis capturing reactiveness to a stimuli. An issue with the valence-arousal model is its difficulty to immediately apply in real-world applications, where applications benefit most when there is a describable entity that can be tied to the output of some computational machinery. On the other hand, discrete emotions like sadness, happiness, and anger, provide us exactly this descriptive power. We note also that recent work in [10] studied discrete emotions for 1,101 user-generated videos, but labels data using 10 annotators following an unspecified "detailed definition of each emotion." Meanwhile, we present baselines on a dataset of animated GIFs annotated by over 2.5M users with soft labels on 17 discrete emotions.

3. GIFGIF DATASET

We gathered data from a website created by Human-Computer Interaction researchers at the MIT Media Lab called GIFGIF[1]. We

[1]http://gifgif.media.mit.edu



GIF-to-GIF Comparison | Animated GIFs | Emotion Scores

Figure 3: GIFGIF Dataset. Left: Screenshot of the GIFGIF interface designed at the MIT Media Lab. Users are asked to vote which GIF best expresses an emotion. Center: Several example animated GIFs from the dataset collected from GIFGIF. Right: Last frame of example GIFs for three emotions along with their emotion scores, shown sorted by their score on a [-1,1] scale where 0 is neutral.

collected 3,858 GIFs on April 29, 2014 from GIFGIF along with their crowdsourced annotations. The GIFs, which have at most 303 frames each, are sourced from a large variety of domains from films, television shows, cartoons/anime, sports, video games, advertisements, user generated content, and user-edited content. The GIFs span a wide range of camera angles, illumination, special effects, humans & non-humans, B/W & color, zooming, resolution, and some original content from as early as the 1930's.

As shown in Figure 3, the website that the MIT researchers designed presents users with a pair of GIFs and asks "Which better expresses *X*?" where *X* is one of 17 emotions: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, happiness, pleasure, pride, relief, sadness, satisfaction, shame and surprise. Users answer with the GIF they perceive expresses the emotion best or select neither. This particular question of "expression" is precisely an effort to capture emotions users perceive in the content rather than what users feel after seeing the GIF.

At the time of our data collection, GIFGIF aggregated over 2.5M user annotations to produce a 17-dimensional vector for each GIF containing a score between 0 and 50 for each emotion where 25 is neutral. The creators of GIFGIF chose to calculate these scores using the TrueSkill algorithm [8], a method originally designed for ranking video game players given the outcome of a game. Here, GIFs replace players and GIFGIF users provide outcomes of GIF pair match-ups. The range $[0, 50]$ comes from heuristic parameters in the TrueSkill algorithm that define a mean performance $\mu = 25$ and uncertainty $\sigma = 25/3$ per player or GIF. In all our experiments and analysis, we normalize this range to [-1,1] for convenience. GIFGIF labels are unique in that each GIF has soft labels for each emotion whereas other datasets [2, 10, 11, 14, 15] have traditionally only had one categorical emotion assignment per image or video.

4. PERCEIVED EMOTION PREDICTION

Given animated GIFs and emotional scores along 17 discrete perceived emotions, we seek to computationally predict these scores in a regression framework and evaluate performance.

4.1 Feature Representations

To better understand what features aid the prediction of viewer perceived emotion, we use four image feature representations, each chosen for their previous use or connection to emotion recognition in visual content. The interested reader is encouraged to see the individual references relevant to each feature for details.

Color histograms: We compute frame-level color histograms in HSV color space for its classical use in vision and affect computing.
Face expression: Facial expressions in GIFs intuitively impact how viewers perceive conveyed emotions. We use a convolutional neural network with top-level one-vs-all SVMs with squared hinge losses which achieved the best validation performance of 69.4% in a public evaluation [16] and represents the current state-of-the art. The training set of [16] consisted of 28,708 48×48 face images over seven emotions: angry, disgust, fear, happy, sad, surprise and neutral. We perform face detection using OpenCV's Haar-like cascade and apply facial expression recognition on the largest face for a 6-D vector of SVM score outputs as a feature.
Image-based aesthetics: Earlier works have shown that emotion has some intrinsic correlation with visual aesthetics [3, 10, 12]. We compute a subset of well-known image-based aesthetic features as described in [3]. GIF frames are divided into 3×3 cells from which cell-level statistics are computed including the dark channel, luminosity, sharpness, symmetry, white balance, colorfulness, color harmony, and eye sensitivity. The normalized area of the dominant object and distances from the dominant object's centroid to grid-line intersections are also computed at the frame-level. Together these form a 149-D feature vector for each frame.
SentiBank: We use a recent mid-level visual representation composed of visual sentiment detectors called SentiBank [4] discussed earlier. The feature consists of a 1,200-D vector of Linear SVMs outputs estimating the originator's intended emotion via "adjective-noun pairs" (ANPs). SentiBank has been shown to work well in other emotion recognition tasks [3, 4, 6, 10] and achieved a F-score for recognizing ANPs of over 0.6 on a controlled test set [4].

4.2 Multi-task Emotion Regression

Traditional independent regression like logistic regression or support vector regression ignores the fact that emotion classes can often be related – for example, the sensation of "surprise" is not necessarily orthogonal to the feeling of "fear." We present a novel application of multi-tasking learning (MTL), specifically, multi-task regression (MTR), to solve this problem. In multi-task learning, our goal is to learn the weight matrix $\mathbf{W}$ comprised of t tasks via the optimization $\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) + \Omega(\mathbf{W})$, where $\mathcal{L}(\mathbf{W})$ is the empirical training loss and $\Omega(\mathbf{W})$ is a regularization encoding task-relatedness. Our "tasks" are the 17 discrete emotion classes.

To model the relationship between emotions, one approach is to constrain regressors of different emotions to share a low dimen-

Table 1: Perceived Emotion Prediction on GIFGIF. The normalized mean square error (nMSE) is reported across 17 emotions over five random repetitions of a 20/80% train/test split for color histogram (CH), face expression (FE), Aesthetics (AS) and SB features with ordinary least squares linear regression (OLS), logistic regression (Logit), and low rank multi-task regression (MTR). Lower nMSE indicates better performance.

	OLS	Logit	MTR
CH	1.7398 ± 0.1868	1.6618 ± 0.2991	1.4641 ± 0.1935
FE	0.8925 ± 0.0036	0.9130 ± 0.0030	0.8955 ± 0.0024
AS	1.0440 ± 0.0133	1.0571 ± 0.0116	1.0361 ± 0.0093
SB	1.5694 ± 0.0614	1.4944 ± 0.0593	2.2901 ± 0.1981

sional subspace. Formally, this means we want a low rank weight matrix $\mathbf{W}$ and need to solve the rank minimization $\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) + \lambda \cdot \text{rank}(\mathbf{W})$. This problem is NP-hard in general [17], and a popular solution is to minimize the trace (or nuclear) norm $\|\cdot\|_*$ instead of the rank function. Trace norm regularization in multi-task learning [19] for a data matrix $\mathbf{X}$ takes the form

$$\min_{\mathbf{W}} \sum_{i=1}^{t} \left\| \mathbf{W}_i^T \mathbf{X}_i - \mathbf{y}_i \right\|_F^2 + \rho_1 \|\mathbf{W}\|_* ,$$

where a least squares loss is used for $\mathcal{L}(\mathbf{W})$. We sample ten equally spaced frames for each GIF, or less for shorter GIFs, and apply the emotion labels from GIFGIF's TrueSkill algorithm as weakly supervised labels $\mathbf{y}$ over each of the frames. We ran our experiments over five random repetitions of a 20/80% train/test ratio, find regularization parameters using cross-validation, and report results on the test set where regression outputs are computed by averaging frame-level scores for GIFs in each emotion. We adopt the normalized mean squared error (nMSE) used in previous studies [1, 5] for our experiments. The nMSE is defined as the mean squared error (MSE) divided by the variance of the target vector and assures that the error is not biased toward models that over or under predict.

We compare our approach to classical linear regression with ordinary least squares, i.e. no weighting, and logistic regression. We note that methods like support vector regression could have been used in place of linear regression, but would still learn emotion models independently. Also, despite the pairwise comparisons in GIFGIF labeling, methods like RankSVM could not be used because only score outputs were exposed, not comparison outcomes.

From Table 1, we see that color histograms expectedly performs poorly since emotion is more complex than can be captured by color. The aesthetics feature, though not designed specifically for emotion recognition, still does encode some information related to the perceptual emotion and achieved the second best nMSE. Surprisingly, the SentiBank feature which is tailored as a mid-level semantic feature for aspects of emotion does not perform well. This may be explained by the fact that there is a cross-domain issue from the training set that SentiBank used to our GIFGIF dataset, and that SentiBank has a fairly conservative F-score of 0.6 for its detectors [4]. Of all features, the face expression feature performs the best; we note that we predict the training label average when no faces are detected. The simple intuition is that humans express and perceive much of their emotions through their faces.

Overall, we consistently observed that the best performing emotion was "happiness" followed by "amusement" for all regressors using face expression features, and the worst performing emotion was also consistently "embarrassment." Inspecting the GIFs in the "embarrassment" category, we found that this emotion was heavily dominated by sequences where a person or cartoon would hide their faces with their hands or another object, or would look down hiding their faces via their pose. In these cases, as one would expect, face detection fails because of occlusions. We believe that the "embarrassment" emotion would benefit most from gesture recognition as many of the occlusions are due to hand movement. The good performance on the "happiness" and "amusement" emotions are unsurprising as both emotions are visually expressed with smiles and laughter. However, due the subtle differences between these two emotions, we also believe that this is also one of the reasons why multi-task learning sometimes performs marginally worse due to ambiguities like this in the model. We believe it maybe important for future multi-task emotion models to also regularize on the similarity and not just the dissimilarity of emotion tasks.

5. CONCLUSIONS

We showed that there are different delineations of emotions and presented a computational approach to predicting viewer perceived emotions on a dataset of animated GIFs. Additionally, we showed that emotions need not be decoupled from each other by presenting a multi-task regression approach for jointly learning over 17 discrete emotions. In the future, we will study the gap between intended, perceived and induced emotion as well as study the temporal patterns of emotion.

Acknowledgments

The first author was supported by the U.S. Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

6. REFERENCES

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. In *Machine Learning*, 2008.

[2] Y. Baveye, J.-N. Bettinelli, E. Dellandrea, L. Chen, and C. Chamaret. A large video data base for computational models of induced emotion. In *ACII*, 2013.

[3] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah. Towards a comprehensive computational model for aesthetic assessment of videos. In *ACM Multimedia*, 2013.

[4] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective-noun pairs. In *ACM Multimedia*, 2013.

[5] J. Chen, J. Zhou, and J. Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *KDD*, 2011.

[6] Y.-Y. Chen, T. Chen, W. H. Hsu, H.-Y. M. Liao, and S.-F. Chang. Predicting viewer affective comments based on image content in social media. In *ACM ICMR*, 2014.

[7] J. Fleureau, P. Guillotel, and I. Orlac. Affective benchmarking of movies based on the physiological responses of a real audience. In *ACII*, 2013.

[8] R. Herbrich, T. Minka, and T. Graepel. TrueSkill™: A Bayesian skill rating system. In *NIPS*, 2006.

[9] W. James. What is an Emotion? *Mind Association*, 1883.

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

[11] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE TAffC*, 2011.

[12] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM Multimedia*, 2010.

[13] R. W. Picard. *Affective Computing*. MIT Press, 1997.

[14] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, 2010.

[15] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimedia database for affect recognition and implicit tagging. *IEEE TAffC*, 2012.

[16] Y. Tang. Deep learning using linear support vector machines. In *ICML*, 2013.

[17] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 1996.

[18] X. Wang, J. Jia, P. Hu, S. Wu, J. Tang, and L. Cai. Understanding the emotional impact of images. In *ACM Multimedia*, 2012.

[19] J. Zhou, J. Chen, and J. Ye. MALSAR: Multi-tAsk Learning via StructurAl Regularization. Technical report, ASU, 2012.