Inferring Visual Persuasion via Body Language, Setting, and Deep Features

Xinyue Huang Adriana Kovashka
Department of Computer Science
University of Pittsburgh
{xinyuh, kovashka}@cs.pitt.edu

Abstract

*The computer vision community has reached a point when it can start considering high-level reasoning tasks such as the "communicative intents" of images, or in what light an image portrays its subject. For example, an image might imply that a politician is competent, trustworthy, or energetic. We explore a variety of features for predicting these communicative intents. We study a number of facial expressions and body poses as cues for the implied nuances of the politician's personality. We also examine how the setting of an image (*e.g. *kitchen or hospital) influences the audience's perception of the portrayed politician. Finally, we improve the performance of an existing approach on this problem, by learning intermediate cues using convolutional neural networks. We show state of the art results on the Visual Persuasion dataset of Joo* et al. *[11].*

1. Introduction

Often, what is interesting about an image and makes it famous is not *what* or *who* is portrayed, but rather *how* the subject is portrayed. For example, the images in the top row of Figure 1 are interesting because of the playful light in which Monroe (a) and Einstein (b) are portrayed, and the deep desperation which stems from the image of the Migrant Mother (c). Further, the images which accompany news articles often add a flavor that the text itself does not convey, and sometimes this flavor carries more weight than the text. For example, a picture which represents a politician in a less than flattering light might be used by the politician's opponents to discredit them in the media (see Figure 1 d,e). The implications in some images, such as Figure 1 f, have even changed the course of wars.[1] The ability of pictures to convey subtle messages is widely used in the media. For example, images of disasters arouse our compassion towards the injured, and photographs of politicians make us believe one thing or another about the photo subjects' qualities and abilities.

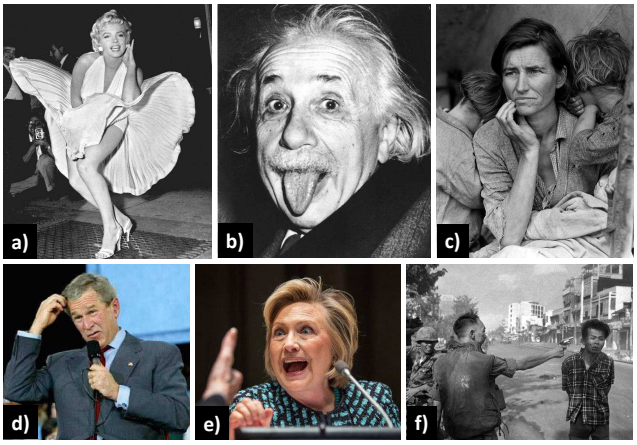


Figure 1. Examples of canonical images whose meaning is not captured by existing computer vision tasks.

If we could develop methods for automatic understanding of how these images portray their subjects and how they affect the viewer, we would enable computers to analyze the visual media around us beyond a listing of the objects found in it, or sentences that describe what takes place in the image. This would mean that machines have come closer to perceiving visual content in a human-like manner. It also has numerous applications. For example, machine understanding of visual intent could allow us to automatically analyze the biases in different media sources, and to curate a sample of news articles about a given event that is balanced in terms of opinions. It would enable us to study how the perception of a politician changed, as chronicled by the media. Finally, it would permit us to analyze how the portrayal of different minorities has progressed historically.

Despite the great value of automatically predicting the subtle messages of images, the computer vision community had until recently never analyzed such high-level tasks. Yet recent progress in computer vision has set the stage for studying these intents. First, we have seen great leaps in the

[1] This image by American photographer Eddie Adams portrays the Vietcong prisoner on the right as entirely helpless, and even though it was taken out of context, it dramatically changed the public opinion of the Vietnam war.

ability to detect and categorize different objects, which is one of the "holy grails" of computer vision [27, 8]. Second, vision researchers have begun to study recognition problems beyond naming objects. Recent work in semantic visual attributes has explored how to be able to describe objects even if the system does not know their names [6], how to perform fine-grained visual search by employing attributes for relevance feedback [15], etc. [5] study the *reasons why* an object category is present, and [24] evaluate the *quality* with which an action is performed. [12, 31] analyze the art styles of paintings and photographs. [22, 21, 33] study the reasons for the successes and failures of vision systems. After these works, it is now time for the community to analyze images at a more subtle, human-like level.

Joo *et al*. were the first to study the "communicative intents" or "visual persuasion" of images [11]. They examine a dataset of eight politicians, and learn to predict whether an image portrays a politician in a positive or negative light, as competent or not, *etc*. Joo *et al*. find that the gestures and facial expressions of politicians, as well as the context and background of the image, can influence the social judgment of those politicians. However, the features that [11] use are limited. We extend their study and improve the intent prediction ability of their system. We explore a wider range of features than [11], including body poses, the setting of the image, and improved facial expression, gesture, and background prediction via features based on convolutional neural nets. We show that capturing a richer set of facial expressions and body poses of the individuals in photographs, as well as the setting in which they are portrayed, improves our ability to automatically predict the intent of an image.

2. Related Work

Visual recognition has come a long way. In recent years, the community has made great progress towards the "holy grail" of computer vision, the ability to accurately detect all categories that a human can describe and localize them in images [29, 7, 27]. We have even begun to tackle problems that bridge visual understanding and language generation, in the form of sentence description [13, 32] and answering questions about images or videos [1, 30]. Most progress in high-level reasoning about images that goes beyond labeling is recent.

One such example of reasoning beyond labels comes in the form of semantic visual attributes. Attributes are high-level semantic properties of objects [6, 17, 23, 16] which allow us to *describe* images even if we cannot *label* or *name* them. Our work shares the spirit of work in attributes, in that we also seek to understand images at a deeper, more descriptive level. Related is also work in fine-grained category recognition [3], which builds models for categories that differ in very fine ways such as bird species, or in understanding actions in terms of how well they are performed, rather than what their category label is [24]. Also in the vein of understanding the visual world at a finer level, particularly with respect to how well we can capture visual information, is work in providing richer supervision to vision systems and debugging them. [5] study the rationales that annotators provide for why a category is present, *i.e*. why a figure skater's shape is good. [22, 21] examine the contribution of data, features and methods for the performance of vision systems. [33] explore the failures of standard computer vision features by inverting them.

Only very recently, the community has begun to tackle some more subtle aspects of images. [12] study the *styles* of artistic photographs. [31] learn to predict *who* took a given photograph, by learning photographer style. [11] seek to understand the "communicative intents" implied in the images, *e.g*. does a photograph portray an individual in a positive or negative light, is the individual shown as competent or not, trustworthy or not, *etc*. To infer these intents, the authors developed 15 types of "syntactical attributes" that capture facial display, gestures, and image background. To experimentally verify how well these syntactic attributes can be used as an image representation in which to learn about communicative intents, the authors developed a database of images of politicians labeled with ground-truth intents in the form of rankings. In our experiments, we study additional features that capture communicative intents.

Note that the explicit or implicit opinions of a text about a given subject have been explored in the natural language processing community in the form of sentiment analysis [20, 34]. Except for [11] who study this problem in the domain of politicians, inferring opinions from *images* has not been explored in computer vision before.

3. Approach

We first present the original features used by [11], and then the novel features with which we experimented. We use the dataset of [11] and aim to learn the 9 communicative intents from that dataset: "favorable", "angry", "happy", "fearful", "competent", "energetic", "comforting", "trustworthy", and "powerful".

3.1. Learning framework

Joo *et al*. [11] defined three kinds of "syntactic attributes"—facial display, gestures, and scene context—for a total of 15 feature dimensions, on top of which communicative intents are learned. They train a linear binary classifier to predict each of the 15 syntactic attributes. Then, they learn a model for the 9 communicative intents, separately for each intent, using the following:

- Descriptors for the training data, which consist of the positive/negative scores for each syntactic attribute;

Feature	Dimensionality
Original syntactic attributes	15
Facial expressions	7
Body poses	30
Scenes	75
PHOG	168
Caffe	4096

Table 1. The features used in our approach.

and

- Ground truth training data in the form of pairs of images (i, j) where image i displays the communicative intent of interest (*e.g.* favorable, competent, or trustworthy) to a greater extent than j does.

Joo *et al*. then learn an SVM ranking model. While [11] use [4]'s ranking SVM implementation, we use [10].

3.2. Original features

For a new image, [11] first predict its 15 syntactic attribute scores.[2] Then they learn a ranking SVM using this syntactic attribute representation, separately for each communicative intent. Note that the dataset provides ground truth bounding boxes for the politician of interest in each image, and the instructions on the dataset note that these boxes can be used at test time, so we use this approach. The features used to learn the 15 syntactical attribute classifiers are as follows:

- **Facial display.** The facial expressions in a photograph can be very telling of a the subject's mental or emotional state, thus they can affect how the subject is perceived (*e.g.* as competent or not). [11] defined four different facial display types: smile/frown, mouth open/closed, eyes-open/closed, and head-down/up. Each image is expected to show a face, and they detect the keypoints on the face using Intraface [35]. For a small patch around each keypoint, they extract HOG features. We follow the same approach as [11], but use [28] to detect facial keypoints since the Intraface software was unavailable for public use. We apply the target person bounding boxes first before extracting the facial keypoints.

- **Gestures.** [11] use seven types of human gestures: hand wave, hand shake, finger-pointing, other-hand-gesture, touching-head, hugging, or none of these. They densely extract SIFT features from the target person bounding box in each image, and compute a 3-level spatial pyramid representation on top of the SIFT features, which is used as the image descriptor for each gesture classifier. We use the same approach. To compute the spatial pyramid representation, we use a vocabulary of K=200 visual words.

- **Scene context.** [11] use four scene context attributes: dark-background, large-crowd, indoor/outdoor scene, and national-flag. They use the same features to capture these attributes as for the gesture attributes, but now extracting features from the whole image. We follow the same approach but, unlike [11], exclude regions containing the bounding box for the person of interest before extracting features, as opposed to using the entire image as in [11].

[2] In our implementation, we use the probabilities of being present that the 15 syntactic attribute classifiers provide.

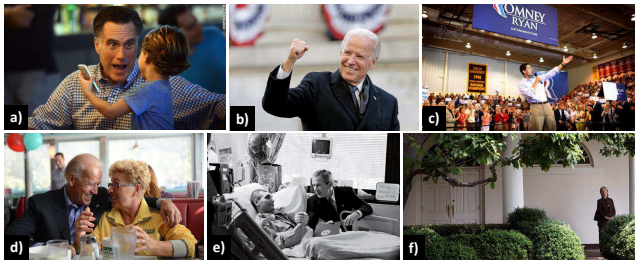


Figure 2. Motivation for the feature types we develop (facial expressions, body poses, and image settings). See text for details.

3.3. Facial expressions

Throughout this work, we use the same learning framework described above to learn communicative intents, but explore new features in addition to the 15 original syntactic attributes. Table 1 summarizes the features.

The facial syntactical attributes used in [11] only capture some facial movements like mouth and eyes being open or closed, but do not capture subtle variations like muscle changes or expressions that do not involve a smile. Consider Figure 2 a. Mitt Romney is portrayed in a positive light in this image because he is able to relate to the child. He has a strong facial expression on his face, but it is difficult to capture this expression with the 4 facial syntactic attributes of [11].

We extend the original set of syntactic attributes with seven new facial expressions corresponding to seven emotion categories from [18]: anger, contempt, disgust, fear, happiness, sadness, and surprise. [18] developed the CK facial expression database, which contains about 324 images, each of which is assigned one expression label. We extracted features for each image using the same method as that used for the facial display attributes. We then trained a linear one-vs-all classifier for each kind of facial expression. For a new test image, we ran each of those classifiers and concatenated their responses to form the 7-dimensional descriptor for the image. These responses are converted to probabilities using Platt's method [25].

3.4. Body poses

Human poses can provide more cues than gestures to predict the communicative intents. Consider Figure 2 b, c. Joe Biden's pose seems to be saying "We can do this", while Paul Ryan's is saying "Here is to the bright future." Neither of these body poses are captured by [11]'s syntactic attributes.

We use [36] to extract the human poses from the images. This method returns the (x, y) locations of 26 different joints, *e.g*. the location of the head, elbows, knees, *etc*. Because there can be multiple human poses detected in one image, we need to combine these multiple poses in a single representation for the image.[3] We experimented with three approaches for doing so, and found that the best strategy is as follows. First, we collect all of the poses detected from the images in the training dataset and use K-means to group them into multiple clusters. We then represent the image as a histogram containing the counts of how many poses of each cluster type were detected in the image. This becomes the pose feature vector for the image. We found that $K = 30$ works best.

3.5. Scene categorization

The backdrop and setting in a photograph contributes to how the subject is perceived. For example, a politician shown in a diner (Figure 2 d) or a hospital (Figure 2 e) might be perceived as being "close to the people." A politician shown in a serene environment outside in nature (Figure 2 f) might be seen as comforting.

Therefore, we represent the image with how well it portrays each of a number of different fine-grained scene categories. We develop classifiers for the 8 outdoor scene categories and 67 indoor scene categories from the datasets of [19] and [26], respectively. [19]'s dataset contains more than 2600 images, and [26]'s more than 15620 images. We extracted GIST as the feature descriptors for learning the 75 (=8+67) scene classifiers.

For each scene classifier, we use the images of the corresponding category label as the positive training data, and all the remaining images in the same dataset ([19] for outdoor scenes and [26] for indoor scenes) as the negative data.

We then train a linear one-vs-all classifier model for each scene category. A novel image is then represented by the concatenation of the responses of the 75 scene category models. Similarly to Section 3.3, the responses are converted to probabilities.

3.6. PHOG

We also extracted Pyramid Histogram of Oriented Gradients descriptors [2] of the whole images as another type of feature descriptor. The dimensionality of this feature is 168. This feature type was intended to represent the images holistically, and to be complementary to our higher-level features capturing facial expressions, body poses, and scenes.

3.7. Caffe

Finally, we experimented with features extracted from a convolutional neural net (CNN). Deep neural networks have achieved state-of-art performance on a number of visual recognition tasks [29, 27, 12], and the responses of various network layers are often used as features. A commonly used CNN toolbox is the Berkeley Vision and Learning Center (BVLC) Caffe package [9]. This framework provides some pertained deep learning models on large-scale image datasets for further recognition tasks. Using the CaffeNet model trained on ILSVRC 2012 [27], we extracted CNN features for the images in the Joo *et al*. [11] dataset from the fc6, fc7, and fc8 layers.

We attempted to use these layers directly as features to predict communicative intents. The fc7 layer performed best among the three, but we found they performed worse than the original 15 syntactic attributes. This justifies the strategy of first computing a low-dimensional set of syntactic attributes, and then predicting intents based on this intermediate representation. Therefore, we experimented with using CNN features to first build classifiers for the 15 syntactic attributes, and use those improved classifiers to predict the 9 intents, which resulted in a boost in performance. For facial display representation, we used the same method as presented in [11]. For syntactical attributes of gestures and context, we use deep learning features to train linear classifiers.

4. Results

We compare the performance of the features described above as a representation in which to learn a ranking SVM for each communicative intent. To compare different methods' outputs, we used Kendall's τ [14] as performance measure, which [11] also used. If the ground truth order for images i and j and intent C is $i > j$, *i.e*. image i has more of intent C than image j does, then a method should produce a higher value for image i than it does for image j in order to be correct for this pair. Kendall's τ then measures the ratio of correctly minus incorrectly ordered pairs, out of all pairs. A higher Kendall's τ indicates better performance. We use the same train/test splits as in [11].

Table 2 shows the result of complementing the original 15 syntactic attributes with different features. The first column shows the performance of the original features in our implementation, which differs from the authors implementation in some ways, as noted in Section 3.2. The next four columns show the addition of the features from Sections

[3]We use all poses because in some cases a persuasive intent is best understood from how two people physically interact.

Intent	Original	Original +Expressions	Original +Pose	Original +Scene	Original +PHOG	Orig (Deep)	Orig (Deep) +Expressions	Orig (Deep) +PHOG
Favorable	25.94%	26.10%	27.64%	31.52%	27.02%	**37.36%**	36.70%	36.64%
Angry	32.98%	32.44%	31.56%	31.66%	33.84%	36.16%	35.10%	**37.34%**
Happy	32.42%	32.18%	30.84%	31.82%	32.88%	39.56%	40.48%	**41.10%**
Fearful	29.28%	29.32%	29.70%	32.42%	31.46%	36.96%	36.60%	**37.44%**
Competent	18.50%	18.42%	16.22%	18.90%	20.16%	**26.34%**	25.98%	26.26%
Energetic	27.38%	27.36%	24.62%	31.62%	28.16%	36.72%	**38.48%**	37.70%
Comforting	18.68%	19.64%	21.46%	23.06%	19.92%	**29.76%**	29.28%	28.70%
Trustworthy	17.60%	16.94%	19.50%	18.52%	16.62%	**30.86%**	30.34%	30.30%
Powerful	20.86%	22.28%	21.26%	26.32%	23.14%	31.64%	**33.86%**	33.70%
MEAN	24.85%	24.96%	24.76%	27.32%	25.92%	33.93%	34.09%	**34.35%**

Table 2. Experimental comparison among combinations of different syntactical attributes. Higher values indicate better performance, using the Kendall's τ metric. Bolded is the best performance in each row.

3.3 to 3.6 to the original syntactic features. We see that on average, expression, scene categorization, and PHOG improve the performance of the original features. The pose features improve the performance for five of the intents, and help the most for the "comforting" one. This could be because whether or not someone is being comforting depends strongly on body language (*e.g.*, is there a hug?), which is captured well by pose features. We also observe that scene categorization features improve the performance of the original features the most. This may be because the original features of [11] only involve four scene attributes so they benefit from the extensive scene categories we detect.

We next examine the performance of using CNN features for learning the original 15 syntactic attributes, which are then used as a representation to learn the communicative intents. In "Orig (Deep)" we see a 37% relative improvement on average, compared to the original features in the first column. Our result of 33.93% is also 13% better than the mean result reported in [11] of around 30%. Thus, the CaffeNet CNN, despite being trained for object recognition, capture cues that are also useful for predicting intent.

Finally, we complement the original syntactic attributes learned with deep features, with our new features. Both facial expressions and PHOG ("Orig (Deep) + Expressions" and "Orig (Deep) + PHOG"), when added to the deep original features, improve their performance.

Figure 3 shows some qualitative examples. We show how the features we develop successfully complement the original syntactic attributes. In particular, we identify pairs of images where the indicated method was able to order these images correctly, unlike the original method of [11]. In the left column, we see examples of the best method, Original (Deep) + PHOG, outperforming the Original method. The correct order, which our method accurately predicts, is that in each pair, the image on the left contains the intent to a larger extent than the image on the right. For example, we can correctly infer that an image showing Barack Obama smiling among a crowd in front of an American flag portrays him as more favorable than one showing him looking down, walking away from the flag. We also see that a picture of him with children, which was a type of image that motivated [11]'s studies, shows him as more favorable than one where children are absent. We also see a similar type of image for Mitt Romney (fifth row). Finally, we see that an image of Paul Ryan in a working environment (last row) portrays him as trustworthy. In all of these cases, the features that the ImageNet-trained CNN learned could be helping us by detecting relevant categories (people, cups, *etc*.)

At the top-right of Figure 3, we see examples of body pose features helping the communicative intent inference. We see that particular body language can help us determine when someone is shown as comforting, and pose features help us capture body language.

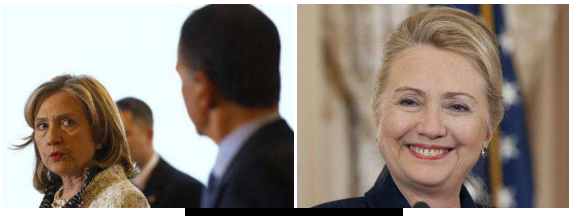
In the remainder of the right-hand column in Figure 3, we see examples of scene features helping our predictions. In the first row, we see that an image of Hilary Clinton in nature (the method is likely unable to infer what monument this is) shows her in a more serene and comforting light than the image showing her in an office-like environment. Similarly, the diner setting in which Joe Biden is shown (second row), and the home and hospital environments shown for George W. Bush (last two rows), portray them in a favorable light. These examples demonstrate how useful capturing image setting is for inferring communicative intents.

5. Conclusion

We develop features that improve a computer system's ability to automatically infer the communicative intents implied in images. Our results shows that besides attributes like facial display, gestures, and background, new features like facial expressions, body poses, and scene categories also prove helpful for predicting the persuasive intents of images. We also demonstrate the value of applying deep

MORE FAVORABLE

MORE FAVORABLE



MORE POWERFUL

MORE COMPETENT

MORE FAVORABLE

MORE TRUSTWORTHY

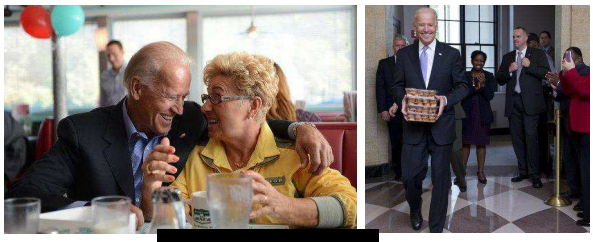
Original(Deep)+PHOG outperforming Original

MORE COMFORTING

MORE COMFORTING

Original+Pose outperforming Original

MORE COMFORTING



MORE FAVORABLE

MORE FAVORABLE

MORE FAVORABLE

Original+Scene outperforming Original

Figure 3. Examples of how our proposed features complement the original 15 syntactic attributes. See text for details.

learning features to predict gestures and background. In our future work, we will examine how our developed features can help us understand the light in which individuals other than politicians are portrayed, and how they can assist us in analyzing news report videos.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and Video retrieval*, pages 401–408. ACM, 2007.

[3] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.

[4] O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with svms. *Information Retrieval*, 13(3):201–215, 2010.

[5] J. Donahue and K. Grauman. Annotator rationales for visual recognition. In *ICCV*, 2011.

[6] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing Objects by Their Attributes. In *CVPR*, 2009.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[8] K. Grauman and B. Leibe. *Visual object recognition*. Number 11. Morgan & Claypool Publishers, 2010.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[10] T. Joachims. Training linear svms in linear time. In *KDD*, 2006.

[11] J. Joo, W. Li, F. F. Steen, and S.-C. Zhu. Visual persuasion: Inferring communicative intents of images. In *CVPR*, 2014.

[12] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. Recognizing image style. In *BMVC*, 2014.

[13] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[14] M. G. Kendall. Rank correlation methods. 1948.

[15] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, 2012.

[16] A. Kovashka, D. Parikh, and K. Grauman. WhittleSearch: Image Search with Relative Attribute Feedback. In *CVPR*, 2012.

[17] C. Lampert, H. Nickisch, and S. Harmeling. Learning to Detect Unseen Object Classes By Between-Class Attribute Transfer. In *CVPR*, 2009.

[18] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, 2010.

[19] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[20] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86. Association for Computational Linguistics, 2002.

[21] D. Parikh and C. L. Zitnick. The role of features, algorithms and data in visual recognition. In *CVPR*, 2010.

[22] D. Parikh and C. L. Zitnick. Finding the weakest link in person detectors. In *CVPR*, 2011.

[23] G. Patterson and J. Hays. SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In *CVPR*, 2012.

[24] H. Pirsiavash, C. Vondrick, and A. Torralba. Assessing the quality of actions. In *ECCV*, 2014.

[25] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[26] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, pages 1–42, April 2015.

[28] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[30] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016.

[31] C. Thomas and A. Kovashka. Who's behind the camera? identifying the authorship of a photograph. In *CVPR*, 2016.

[32] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[33] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. In *ICCV*, 2013.

[34] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT*, pages 347–354. Association for Computational Linguistics, 2005.

[35] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[36] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.