Visual Persuasion: Inferring Communicative Intents of Images

Jungseock Joo[1], Weixin Li[1], Francis F. Steen[2], and Song-Chun Zhu[1]

[1]Center for Vision, Cognition, Learning and Art,
Departments of Computer Science and Statistics, UCLA
[2]Department of Communication Studies, UCLA
{joo@cs, steen@commstds, sczhu@stat}.ucla.edu

Abstract

In this paper we introduce the novel problem of understanding visual persuasion. Modern mass media make extensive use of images to persuade people to make commercial and political decisions. These effects and techniques are widely studied in the social sciences, but behavioral studies do not scale to massive datasets. Computer vision has made great strides in building syntactical representations of images, such as detection and identification of objects. However, the pervasive use of images for communicative purposes has been largely ignored. We extend the significant advances in syntactic analysis in computer vision to the higher-level challenge of understanding the underlying communicative intent implied in images. We begin by identifying nine dimensions of persuasive intent latent in images of politicians, such as "socially dominant," "energetic," and "trustworthy," and propose a hierarchical model that builds on the layer of syntactical attributes, such as "smile" and "waving hand," to predict the intents presented in the images. To facilitate progress, we introduce a new dataset of 1,124 images of politicians labeled with ground-truth intents in the form of rankings. This study demonstrates that a systematic focus on visual persuasion opens up the field of computer vision to a new class of investigations around mediated images, intersecting with media analysis, psychology, and political communication.

1. Introduction

Persuasion is a core function of communication, aimed at influencing audience beliefs, desires, and actions. **Visual persuasion** leverages sophisticated technologies of image and movie production to achieve its effects. The examples in Fig. 1. (a) are designed to convey social judgments: that Obama is an inferior candidate to Romney, that a Mac is more user friendly than a PC, and that Hitler is kind and trustworthy. These claims are not stated verbally, but rely on routine visual inferences.

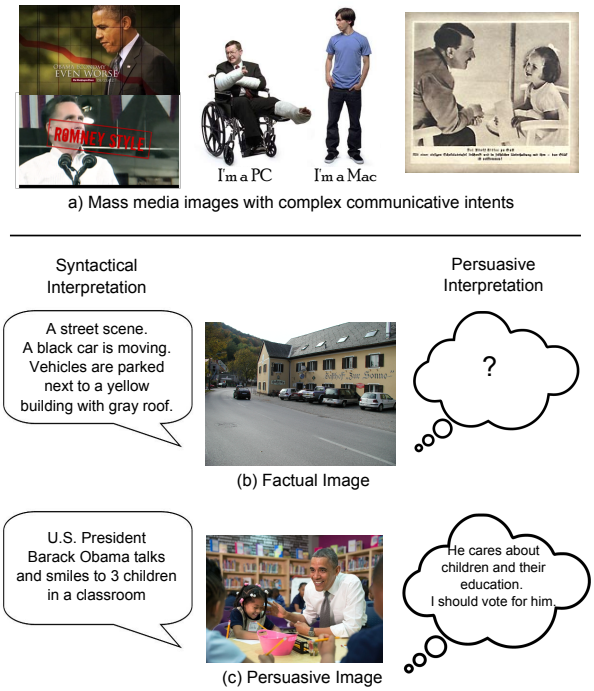

Figure 1. (a) A persuasive image has an underlying intention to persuade the viewer by its visuals and is widely used in mass media, such as TV news, advertisement, and political campaigns. For example, it is a classical visual rhetoric to show politicians interacting with kids, arguing that they are dependable and warm. (b) Existing approaches in computer vision lead to syntactical understanding of images to explain the scene and the objects without inferring the intents of images, which is absent in factual images in usual benchmark datasets. (c) Our paper is aimed at understanding the underlying intents of persuasive images.

Visual argumentation is widely used in television news and advertisements to generate predictable social judgments. Why did President Obama post Fig. 1. (c) on his White House Blog? The image contains no policy-relevant

information. It does, however, lend itself to generating a set of inferences about Obama's character: that he loves children, that he is caring, and that he can be trusted with making the right decisions in education. Such inferences are politically extremely valuable for a politician, and are hard to convey verbally.

Examining the image in more detail, one can notice it contains a suite of syntactical components to compose its intent: the protective gesture, steady gaze, welcoming smile, and the child smiling. Audiences see these elements and make judgments as if they were present, yet what the image shows is the result of professional photographers composing and selecting these elements in order to create a specific impression. Because we believe our own eyes, but know well that people are manipulative, we tend to be verbally skeptical and visually gullible.

In this study, we examine nine different trait dimensions in order to characterize the **communicative intents** of images. To infer these dimensions, we exploit 15 types of syntactical features – facial attributes, gestures, and scene contexts that construct the communicative intent. Computer vision research has made remarkable progress in addressing syntactical problems; we extend these techniques to understand and predict large-scale patterns in the higher-level persuasive messages that images in the media routinely convey. In summary, this study addresses the following research questions:

i. We define a novel problem, to infer the communicative intents from images, in a computational framework. We identify the dimensions of intent in persuasive images and describe how they can be inferred from syntactical features. The complete list of intents is presented in Fig. 2.

ii. We present a new dataset to study visual persuasion. It contains 1,124 images of 8 U.S. politicians annotated with the persuasive intents of 9 types as well as syntactical features of 15 types.

iii. Finally, to verify the impact of visual persuasion in mass media, we present a quantitative result in a case study that reveals a strong correlation between the visual rendition of the U.S. President in mass media and public opinion toward him.

2. Related Work

Our paper is related to studies in computer vision on human attribute recognition [8, 22, 16, 13], such as gender, race, or facial expression recognition. However, communicative intents are distinct from traditional human attributes in two important ways. First, intents focus on judgment rather than surface feature. We deploy syntactic interpretation to leverage surface features as intermediate representations. In our analysis, persuasive intents

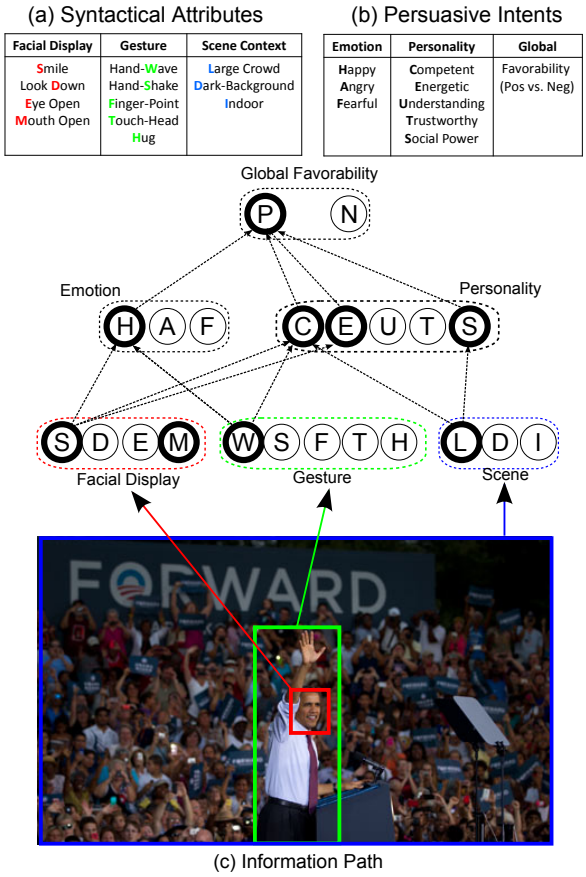

Figure 2. The set of (a) syntactical attributes and (b) communicative intents defined in this study. (c) Illustrative hierarchy of intention inference. The image is first interpreted syntactically. Second, the syntactical representation is transformed to infer the communicative intents. Third, communicative intents in the form of emotional characteristics and personality traits are used to assess global favorability.

are not directly observable, but inferred from complex patterns involving multiple image evidence. Second, the syntactical feature can have specific social semantics beyond its surface, narrative, first-order meaning. For example, a "hand wave" can mean "competence" or "popularity," while "touching face" can imply "trouble". We seek to systematically identify these underlying implicatures [10], or hidden semantics, of the syntactic attributes, which have not been considered in the prior works. The distinction between syntactic features and communicative intents in visual communication parallels the distinction between literal message meaning and communicative intention in pragmatic theories of language [1].

Researchers in political science and mass media have examined audiences' emotional and cognitive responses to televised images of political leaders [17, 24] and studied the media's selective use of images for persuasive purposes [2, 20, 21, 9]. This body of work has reported a series of correlations between politicians' appearance on media and

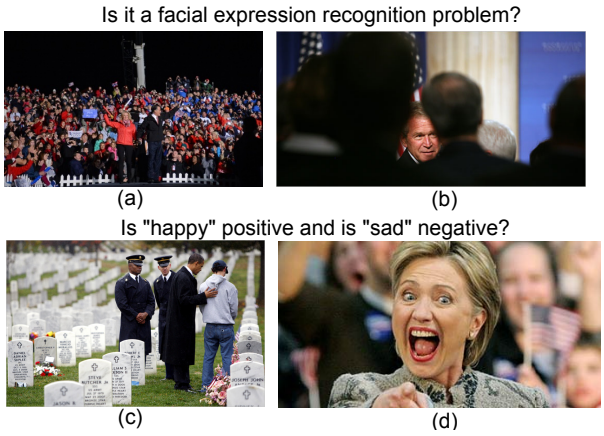

Figure 3. (a-b) Communicative intents can be inferred from non-facial cues such as gestures (*e.g.*, hand wave), or scene context (*e.g.*, large crowds). (c-d) The intents cannot be understood coherently by a single dimensional approach such as polarized sentiment analysis. The annotators found the image (c) "sad" but also "comforting", so they believe the image shows "positive favorability" toward the target, whereas the exact opposite observations are found in the image (d).

electoral success. While their results are interesting, they are restricted by manual analysis to a small number of images or television news shows.

3. Representation

In this section, we identify the dimensions of communicative intents as well as the syntactical features used to predict the intents. Fig. 2 highlights the overall representation of our model where the layers of hierarchy are defined according to the levels of abstraction. An image can be first interpreted at the syntactic level by many types of 'detectors' or 'classifiers', such as a smile classifier. The outputs of the syntactic solvers are used to infer a set of communicative intents, which we aggregate to measure the global favorability of the image, positive or negative.

3.1. Dimensions of Communicative Intents

What are the intentional dimensions we can use to specify the perception that the image conveys to viewers? One can imagine there exist thousands of different types of emotions or feelings that we can attribute to the main target of image. To simplify the problem, we choose a series of quantized dimensions of the communicative intents as follows.

i. *Emotional Traits*. Ekman, in his influential early work in psychology [6], has argued the human has 6 elementary and 'universally interpretable' emotions. Masters *et al*. [18], have further defined 10 emotional traits in their study on analyzing facial displays of politicians. Since this set covers a broad basis of human emotions that can apply to general domains including politics, we adopt these 10 emotional traits in our study.

ii. *Personality Traits and Values*. While the emotional dimensions focus on the subjects' own emotions, what one can perceive from an image and what an image is ultimately intended to convey requires a broader and more general scope. For instance, 'understanding (the others)' is an important requirement to politicians and many social figures and persuasive images often achieve this concept by showing the targets interacting with general public, the handicapped, or children. These traits describe more essential and canonical characteristics of people than the spontaneous captures of emotions. Therefore, we consider six additional dimensions in this category which are particularly important to the politicians.

iii. *Overall Favorability*. Finally, we also define the overall favorability of a target person as an integrated measure. This can be viewed as a summarized polarity, *i.e.*, *positive* or *negative*.

Given initial dimensions, we conducted a preliminary test on a subset of our dataset to merge redundant dimensions, *i.e.*, multiple dimensions producing very similar ratings, and obtained 9 dimensions as follows: angry, happy, fearful, competent, energetic, comforting, trustworthy, socially dominant, overall favorable.

3.2. Syntactical Attributes

There are many types of visual cues from which we can predict the communicative intents. Some cues may be retrieved from the target person, especially from the face and gesture, and other cues from scene context and background. These cues can be viewed as the syntactical interpretation of image and they can collectively signal toward particular intents at a higher and more abstract level. Specifically, we use 15 syntactical attributes categorized into 3 groups as follows.

Facial Display. Psychological studies of emotion discrimination have long focused on the sophisticated human ability to read faces [6]. From the computational perspective, it can be said that the facial region provides the most distinguishable cues to classify attributes [22]. We use 4 different facial display types as follows: smile/frown, mouth-open/closed, eyes-open/closed, and head-down/up. To recognize each display type, we first detect the face and localize the facial keypoints (the centers of both eyes, mouth, and the whole face) by Intraface [25], which is a publicly available software. From each keypoint, we extract HoG feature of 4×4 cells at 3 different scales and then train a linear SVM for each type.

Body Cues - Gestures. Body language often provides a useful channel in non-verbal communication and similarly, certain gestures or actions can deliver or form the sentiments about the target person. An example is "hand wave"

Figure 4. An example question given to the annotators. Given a pair of images, each annotator can respond to judge which image has greater intention to emphasize a certain emotion or personality in the given dimension.

by politicians, which can be viewed as an positive action to show an engagement between the politician and the electorate [4]. "Hugging" is another example which asserts a similar function, possibly with a greater strength as it involves a physical contact. We define 7 types of human gestures frequently used in the domain of political news as follows: hand wave, hand shake, finger-pointing, other-hand-gesture, touching-head, hugging, or none of these (normal).

Gesture or action recognition from images is a difficult problem due to pose variation and occlusion. Moreover, we are interested in distinguishing subtle differences such as hand wave and finger-pointing, which might make the existing pose-driven approaches ineffective. However, it goes beyond the scope of this paper to address such challenges in detail. We adopt a simple method of 3-level spatial pyramids with densely sampled SIFT features encoded by the dictionary learned by K-means clustering, which has been proposed for action recognition in the recent literature [5].

Scene Context. Finally, persuasive intents can be also inferred from the image background as it provides contexts and situations the target is facing. For example, a large group of supporters can imply strong popularity of the target and a dark background may hint at an uncertain future of the target ("doomed"). We use 4 binary scene context features as follows: dark-background, large-crowd, indoor/outdoor scene, and national-flag. To classify each scene context type, we use the same method used for gesture type recognition, discussed above.

Once all feature type values are obtained, each image, I, can be represented by a 17 dimensional response vector at syntactical level. We denote the response vector by $\mathbf{f}(\mathrm{I}) = [f_1(\mathrm{I}), f_2(\mathrm{I}), ..., f_p(\mathrm{I})] \in \mathbb{R}^p$, where the subscript specifies the feature type and p, the number of dimensions, equals 17 in this paper.

4. Learning to Rank Communicative Intents

4.1. Developing a scale

Relative vs. Absolute Assessment. Our goal is supervised learning of mapping functions from the images to a series of communicative intents (Sec. 3.1), which requires supervision as ground-truth annotation. In many syntactical problems, the outputs to predict are defined as binary or discrete variables with predefined categories and in this case the absolute assessment of images from the annotators is suitable (*i.e.*, fact-checking). In contrast, we require our annotators to report their own subjective judgments based on their perceptions and thus, we cannot simply ask them to assign an absolute score to each image since they do not share the same reference scale. Moreover, the magnitude of signals of the intents may be very subtle, which makes absolute assessment even harder.

Fortunately, we observe that communicative intents, although ambiguous when evaluating each image individually, can be much better grounded on *relative* scale such that the valence of an image in a perceptual dimension emerges from the comparison against the other images. Similar motivations can be found in literatures of information retrieval and computer vision [12, 23, 15]. We therefore present a pair of two images at a time and ask the annotators to choose which image depicts the higher degree of given trait, *e.g.*, "Which one looks more positive?". Fig. 4 shows an example question and an image pair.

We can now introduce the notation for the intents. Given a pair of images $(\mathrm{I}_i, \mathrm{I}_j)$, $\mathrm{I}_i \succ \mathrm{I}_j$ indicates the image I_i has a proceeding order to I_j. Since each response only provides a pair-wise order, we aggregate all responses to construct the global ranking order using HodgeRank [11] which arranges the entire image set in one sequential order. One motivation of having the global ranking is to deal with inconsistent pair-wise annotations ($A \succ B, B \succ C, C \succ A$).

Individual vs. Universal Rank. In this study, we do not consider the difference between individuals but solely focus on the difference between the different photographs of the same person. The purpose of this treatment is two-fold. First, we want to rule out the annotators' personal preference on certain politicians that may be based on historical records or stereotypes (gender or age). Evidently, these are not observable from the images. Second, our overall study is aimed at understanding the editorial intents conferred on images by altering specific features. However, the individual factors such as gender or their original facial appearance cannot be altered by editorial tastes as they are constant.

Therefore, we restrict the pair-wise comparisons and the global rank to be obtained from the same subject in order to systemically exclude the individuated factors and existing political preference of human annotators in our study. For example, we do not attempt to compare an image of "Obama" with another image of "Romney". Our treatment is exactly *orthogonal* to that of [23] in which all image instances of one person share the same attribute value. This is because they model person-specific and image-invariant attributes such as "big-lips", while we seek the image-specific attribute values.

4.2. Model

Our model to predict the intents - the upper structure of the overall hierarchy in Fig. 2 - builds on the framework of Ranking SVM [12]. While the binary SVM attempts to maximize the margin between the examples (suppert vectors) of two classes, the ranking SVM maximizes the pair-wise margins in the order specified by the training set. For each dimension of persuasive intents, we are given N training images and their global ranking order, $D = \{(i,j)|\mathrm{I}_i \succ \mathrm{I}_j\}_{i,j=1}^{N}$. We first obtain the syntactical feature vector (Sec. 3.2), $\mathbf{f}(\mathrm{I}) \in \mathbb{R}^p$, for each image. Next, our goal is to learn a linear ranking function, $r(\mathrm{I}) = \langle \mathbf{w}, \mathbf{f}(\mathrm{I}) \rangle$, with the following objective:

$$\begin{aligned} \text{minimize :} \quad & \frac{1}{2}||\mathbf{w}||_2^2 + C\sum \xi_{i,j} \\ \text{subject to :} \quad & \mathbf{w}^\top \mathbf{f}(\mathrm{I_i}) \geq \mathbf{w}^\top \mathbf{f}(\mathrm{I_j}) + 1 - \xi_{i,j}, \\ & \xi_{i,j} \geq 0, \forall (i,j) \in D, \end{aligned} \tag{1}$$

introducing a non-negative slack variable, $\xi_{i,j}$, for every pair in D. C controls the trade-off between training error and margin maximization. We use the implementation of [3] to solve this optimization problem. Finally, since we assume the global favorability can be inferred on the basis of the other types of intentions, we train its ranking function with the outputs from the second layer as well as from syntactical features.

Universal Model. If we train a separate model for each person, it is likely to produce better prediction performance, but at cost of higher complexity and limited scalability. We train one unified model that can address universal characteristics shared by the group of different politicians. Hence, the ranking order set, D, contains the example pairs of all people but does not have any pair of images of different individuals. This should not be confused with individual vs. universal rank discussed above; the annotation and the evaluation still follow the individual protocol but we learn one shared model.

5. Experiments

5.1. Dataset: Persuasive Portraits of Politicians

We present a new dataset of 1,124 images of the politicians with the labeled communicative intents and syntactical features. Fig. 5 shows a few examples. While our methodology is applicable to general domains, we specifically choose the political domain in our dataset because this is where the media would have the biggest intention to persuade the audience. Also, we can easily find hundreds of different photographs of the same politician. This unique property enables the media to *deliberately* select which image to present according to their own editorial tastes and news contexts.

Specifically, we chose 8 U.S. high-profile politicians[1] whose has frequently appeared in the main stream media and collected their photographs from many news outlets online. 10 undergraduate and graduate students participated in annotation process. As discussed earlier, we let the annotators rate the ordinal values by comparison from pairs of photographs of the same politician, from which we recovered the global rank. We also labeled the syntactical attributes used in Sec. 3.2 and a bounding box to specify the main target of each image.

Annotator Agreement. It is important to verify how much consensus the annotators had in evaluating intents. We measured the correlations among the annotators by retrieving an independent global ranking order for each annotator and obtaining correlation coefficients between every two ranking orders from different annotators, which ensured a high degree of agreement (0.647).

5.2. Predicting Communicative Intents

Baseline. Since there are no existing methods developed for our problem, we trained baseline classifiers which take the low-level image features and directly output the intents. Our baseline classifiers adopt 3-level SPM with densely sampled SIFT features, which is the same method that we use for recognizing gesture types.

Measure. We use Kendall's τ [14] (or Kendall rank correlation coefficient) as performance measure, which is common in ranking evaluation [12]. Given two global rankings (one from ground-truth and the other from prediction in our case), it measures how similar or dissimilar they are by counting pair-wise consistencies for all pairs. It is simply defined as follow:

$$\tau = \frac{(\#\text{ of concordant pairs}) - (\#\text{ of discordant pairs})}{\#\text{ of all possible pairs}}.$$

A concordant pair means two examples in the pair are observed in the same order in both ranking sequences. If two rankings are identical, Kendalls' tau equals to one. If they are reversed, it is negative one. Since we separate the ratings for different politicians, we measure the accuracy for each person and use the average value.

Given this baseline and measure, we evaluate how well the learned model can predict the communicative intents of images. Fig. 6. shows the results. First, one can see the dominant effect of the facial displays in the emotional dimensions. Indeed, a face provides an instant delivery of emotional states and some other dimensions such as "trustworthy" can be also perceived and inferred from face. At the same time, we also observe the other cues, such as gesture types, can better predict certain dimensions such as "comforting" or "socially dominant". When do we feel that

[1] Barack Obama(199), George W. Bush(174), Mitt Romney(154), Hillary Clinton(152), John McCain(153), Joe Biden(121), Paul Ryan(82), and Sarah Palin(89), (# of images of each person).

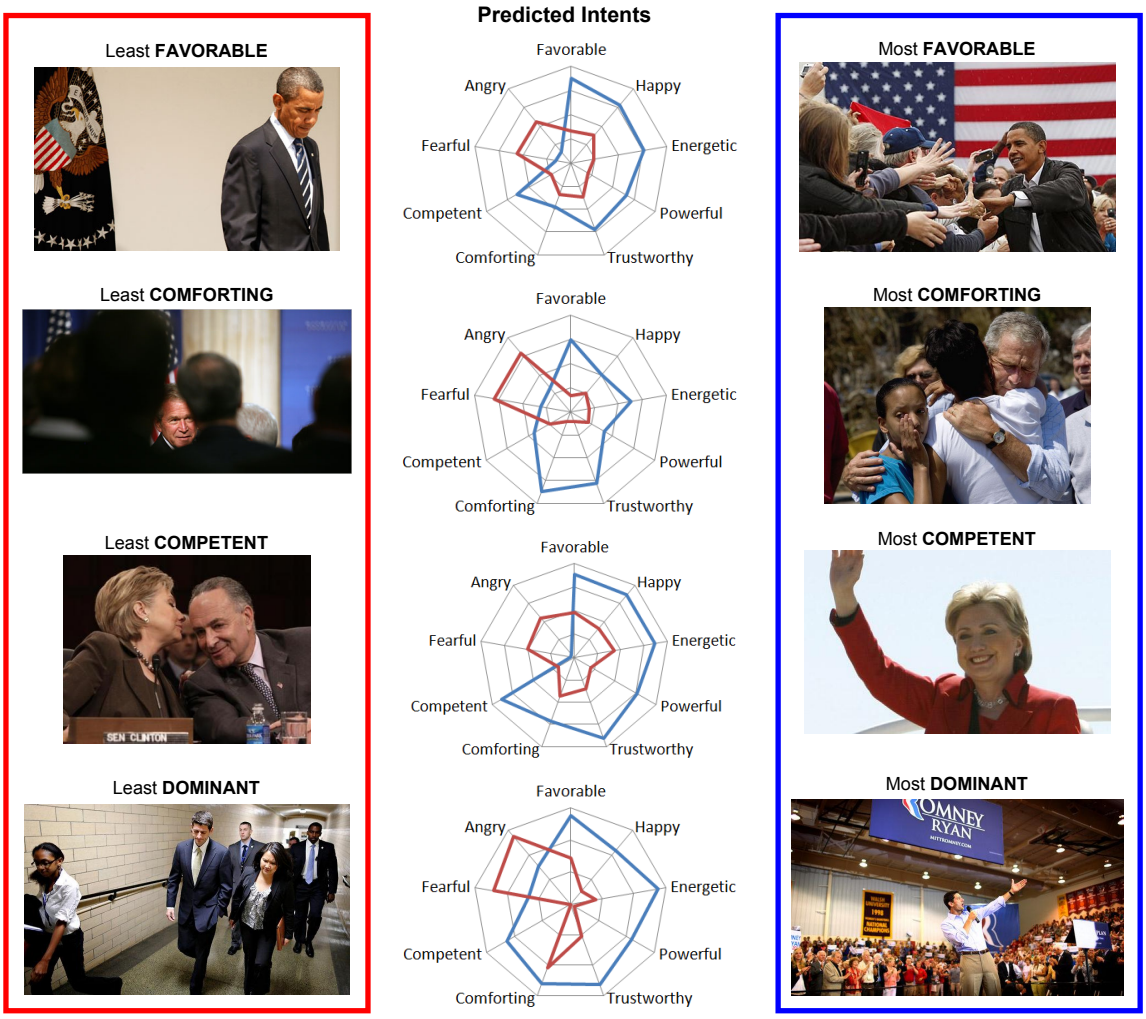

Figure 5. Example images in our dataset and their intents inferred by our model. The images of the right side are "the most" examples in given dimensions and plotted by "blue" curves, whereas the left side are "the least" examples plotted by "red" curves.

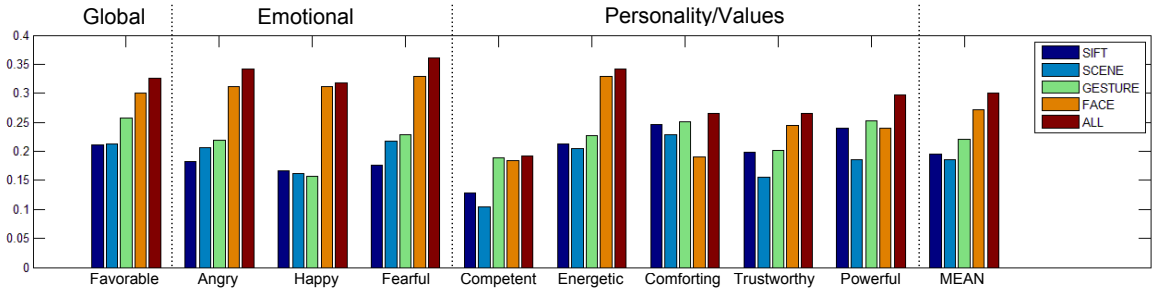

Figure 6. Intents prediction performance evaluation measured by Kendall's tau. For all dimensions, the full approach that exploits all three types of syntactical features yields the best result. In addition, the facial display type outperforms the other cues on the emotional dimensions while the gesture type is more discriminative for 3 among 5 dimensions of personality traits and values.

someone is comforting from the image? To quantitatively answer this question, we further investigate what are the particular causes to invoke these sentiments. Fig. 8. shows the correlation coefficient matrix between the ground-truth syntactic features and the intent annotation. From this matrix, one can say, for example, perception of competence arises from combination of facial displays (smile), gestures (hand wave, hand shake), and scene context (large-crowd).

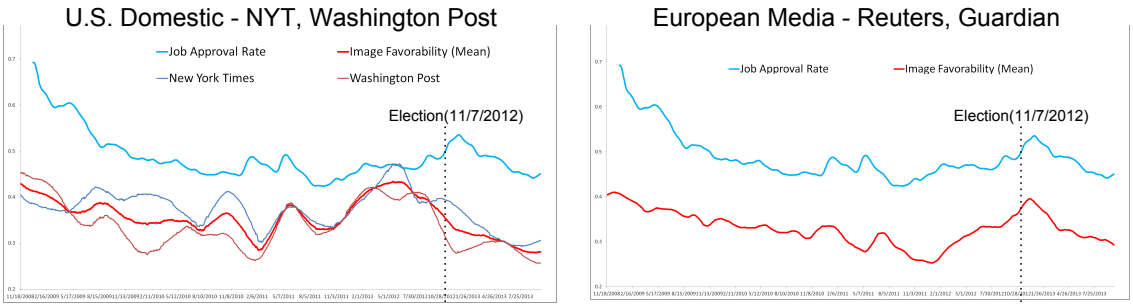

Figure 7. Contrastive media coverage: agenda-setting and mirroring. The correlation between the computed image favorability and public opinion can reveal different motivations of the media outlets.

Syntactical Feature Recognition. Table. 1 reports the accuracy of our model to recognize syntactical attributes. In particular, some fine gesture types are very difficult to distinguish: finger-pointing vs. other-hand-gesture (*e.g.*, fist or v-shape). If we replace the intermediate-level response vector by ground-truth attribute annotations, the intent prediction accuracy improves up to 0.48 (compared to 0.30 in the fully automated model), suggesting room to improve.

5.3. Media and Public Opinion

Now we consider the visual persuasion in the real-world news stories in which the 3rd party newsmaker (*e.g.*, Reuters) describes the target person (*e.g.*, US president). In most cases, the news stories are delivered to the audience through the verbal (text) and optionally the visual (image), which complement each other: the verbal elaborates the details of an event and the visual spotlights the key aspects to consider.

Table 1. Accuracy of Syntactical Feature Recognition.

Category	Facial	Scene	Gesture
AP	.744	.658	.351
Frequency	.392	.267	.143

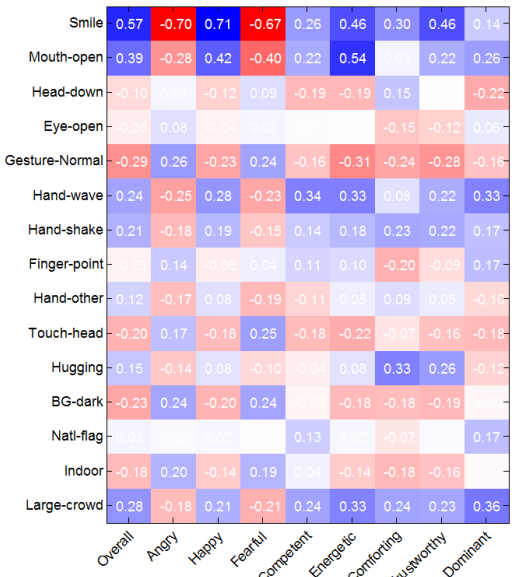

Figure 8. Correlation between the ground-truth syntactic features and the intents.

In this case study, we examine the temporal behavior of the media in framing a particular subject by presenting the visuals and its relation to the public opinion about the subject. In particular, we choose Barack Obama, the 44th and the current President of the U.S., as our main target because he has gained the biggest attention from the U.S. domestic media as well as international media over past 5 years. Moreover, there exist many types of polls to measure the public opinion about him during this period.

We processed the data as follows. First, we crawled all newspaper articles that mention his name from 4 sources - the New York Times, Washington Post, Reuters, and The Guardian - from 01/01/2008 to 09/30/2013. Then we discarded the articles without any photographs and the ones whose photograph captions do not specify the name of the target. For each image, we obtained the syntactical features as follows. Since there is no bounding box available, we first identified Barack Obama's face by the same approach taken for recognizing the facial display types (Sec. 3.2). Once the system recognizes him, it infers the facial display types from the recognized face. For gesture type, we used deformable part model [7] to detect the person bounding box and continued recognizing gestures. The scene context feature can be inferred irrespective of prescribed bounding box. The remaining stages are the same as our model discussed and finally the system outputs the overall favorability of each image. At a data point (a particular date), we smoothed the prediction scores of images of the articles around the date within 1-month time window. To compare this computed statistics against, we obtained the presidential job approval rates from Huffington Post, aggregating 2,356 polls from 86 pollsters.

Figure 7. shows the temporal evolutions of the media's presentation (red) and the public opinion (blue). We see that the detected overall image favorability in the images from foreign sources (Reuters and The Guardian) correlates more closely with the aggregate opinion polls than the images from the politically most influential US newspapers (New York Times and Washington Post). The closer correlation ($\rho = 0.756$) in the foreign sources suggests that these

news outlets are passively **mirroring** the ups and downs of Obama's facial expressions in public events. The graph shows him peaking at the previous election, dropping down afterwards, and then rising back up to peak some time after his re-election in November 2012, clearly showing the dramatic effects of the elections and their immediate aftermath.

In contrast, the prominent domestic newspapers show less correlation overall ($\rho = 0.362$), and significantly a sharp negative correlation with opinion polls from the reeletion. Images used in these news media appear to take the lead in showing less favorable images of the President, foreshadowing the drop in popularity that follows some weeks after the election. Media scholars have argued that the mass media play an important role in setting the **public agenda** [19]. These results are consistent with a large body of work in media studies and political communication, using data sources and methodologies that have previously been unavailable.

6. Conclusion

This study aims to demonstrate that a systematic examination of communicative intents can yield new insights into the meaning and persuasive impact of images, which goes far beyond traditional classification on surface features. We contribute a new dataset of political images, and show how to build on an advanced syntactical analysis, and infer multiple dimensions of persuasive intents. Finally, we show that the resulting favorability judgments correlate in informative ways with independent measures of public opinion, proposing the contrasting media behaviors of agenda-setting and mirroring. By engaging with the ubiquitous use of visual images in the mass media, computer vision can make unique new contributions to an emerging field.

Acknowledgements. This work was supported by NSF CNS 1028381, ONR MURI N00014-10-1-0933, and NSF IIS 1018751.

References

[1] J. L. Austin. *How to Do Things with Words*. Clarendon, 1962.

[2] K. G. Barnhurst and C. A. Steele. Image-bite news the visual coverage of elections on US television, 1968-1992. *The Harvard International Journal of Press/Politics*, 2(1):40–58, 1997.

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

[4] P. D. Cherulnik, K. A. Donley, T. S. R. Wiewel, and S. R. Miller. Charisma is contagious: The effect of leaders' charisma on observers' affect. *Journal of Applied Social Psychology*, 31(10):2149–2159, 2001.

[5] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010.

[6] P. Ekman and et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.

[8] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *NIPS*, 1990.

[9] M. E. Grabe and E. P. Bucy. *Image Bite Politics: News and the visual framing of elections*. Oxford University Press, 2009.

[10] H. P. Grice. *Logic and conversation*. Dickenson, 1967.

[11] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye. Statistical ranking and combinatorial hodge theory. *Math. Program.*, 2011.

[12] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD*, 2002.

[13] J. Joo, S. Wang, and S.-C. Zhu. Human attribute recognition by rich appearance dictionary. In *ICCV*, pages 721–728, 2013.

[14] M. Kendall. *Rank Correlation Methods*. Griffin, 1948.

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

[16] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. *TPAMI*, October 2011.

[17] J. T. Lanzetta, D. G. Sullivan, R. D. Masters, and G. J. McHugo. Emotional and cognitive responses to televised images of political leaders. *Mass media and political thought: An information-processing approach*, pages 85–116, 1985.

[18] R. Masters, D. Sullivan, J. Lanzetta, G. Mchugo, and B. Englis. The facial displays of leaders: Toward an ethology of human politics. *Journal of Social and Biological Systems*, 9(4):319–343, Oct. 1986.

[19] M. E. McCombs and D. L. Shaw. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187, 1972.

[20] P. Messaris and L. Abraham. The role of images in framing news stories. In S. Reese, O. Gandy, and A. Grant, editors, *Framing Public Life: Perspectives on Media and Our Understanding of the Social World*, Routledge Communication Series. Taylor & Francis, 2001.

[21] J. E. Newhagen. The role of meaning construction in the process of persuasion for viewers of television images. In M. P. James Price Dillard, editor, *The Persuasion Handbook: Developments in Theory and Practice*. 2002.

[22] C. Padgett and G. W. Cottrell. Representing face images for emotion classification. In *NIPS*, 1997.

[23] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[24] D. G. Sullivan and R. D. Masters. "happy warriors": Leaders' facial displays, viewers' emotions, and political support. *American Journal of Political Science*, 1988.

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