# Estimating Desired Actions Stimulated by Annotated Images

Bei LIU†, Makoto P. KATO†, and Katsumi TANAKA†

† Department of Social Informatics, Kyoto University
Yoshida-honmach, Sakyo-ku, Kyoto
606–8501 Japan
E-mail: †{liubei,kato,tanaka}@dl.kuis.kyoto-u.ac.jp

**Abstract** This paper addresses the problem of finding desired actions motivated from images with annotation. Many images in the Web are created to promote, trigger, or motivate for people to take some actions after viewing those images (called Call-to-Action, CTA), such as advertisement images. We decompose the problem of learning the relationship between images and desired actions into two sub-problems for which plenty of text data are possibly available for learning. The first sub-problem is to estimate objects and impressions from images, while the second sub-problem is to estimate desired actions from the combination of objects and impressions. We conducted experiments to demonstrate the performance of our desired action prediction, and showed that the problem decomposition could improve the desired action prediction especially when a limited number of manually labeled images were available.

**Key words** Image, desired action, object, impression

## 1. Introduction

We are surrounded by a huge number of images online and offline nowadays. Images have become a way of expression and a resource of inspiration. What we can perceive from images has been widely researched, including object recognition and action detection. Recently, how images could affect us human becomes to attract attentions while retain undocumented, especially after the popularity of image sharing website, such as Pinterest. According to [16], 85.6% of respondents agreed or strongly agreed that they use Pinterest to find "something I might be doing later" and 86.7% agreed to use Pinterest to look for "inspiration for something I will be doing soon". We have to admit that we are affected by those images we look at every day, or those images are encouraging us to take further actions. Some images are created to promote, trigger, or motivate people to take actions on purpose, such as advertisement images. Some affected us unconsciously, like images shared by friends in social network. Most of us have experience that after looking at an image about delicious food, we will feel hungry, even if we just had dinner two hours ago. Another typical example is when we see very beautiful images shared by friends about their trip, we would feel like to go the same place. We can even assume that images taken by friends have more attraction than advertisement images about the same place, due to people's trust in friends and expectation to share similar experience with friends.


Figure 1 User's reaction to an image with fresh ingredients.

What happens when we look at something? This question has been researched a lot in the field of psychology. A commonly know theory is "ABC of psychology" [15] which involves affect, behavior and cognition. However, there is no accurate saying about their relationship. In this research, we are not going further to discuss them. Instead, we will use some concepts to help with the explanation. Our first interaction with stimuli is perception, which is out of our awareness. In Goldstein's book about sensation and perception [4], he claimed that perception is the result of complex "behind the scenes" processes. In the case of image viewing, perceiving information in the images, which is called perception, is conscious sensory experience. Then our past knowledge leads us to place objects in a category, such as recognizing car in the image as "car". Having the concept been categorized in our mind, we might feel like to move our head to focus more on the image, which is defined as actions. Another key part when we look at an image is affect, which is the experience of feeling or emotion. For example, an im-

age of scary snake will make us feel upset while images of beautiful scenery will please us.

In this research, we are talking about a further concept in affect than emotion and arousal, which we call it desired action. For example, image in Figure 1 is about somebody dealing with fresh ingredient, cutting mushroom. Based on our general knowledge, we can imagine that he/she is preparing for cooking. The freshness of ingredients in the image make viewers feel pleased and happy. Moreover, viewers may feel like to eat vegetables or cook by themselves. In this case, eating vegetables and cooking is the desired action that are stimulated by the image.

In this work, we are trying to find what desired action could be stimulated by an image and estimate to what extent the desired action is stimulated. However, since the appearance of machine, semantic gap has been a significant problem for computers, not to mention understanding people's reaction to a certain type of information by computers automatically. In this paper, we would like to introduce our approach to solve this problem by decomposing it into two sub-problems, for which external resources and information can be utilized. The first sub-problem is to extract objects and impressions from annotated images, and the second sub-problem is to estimate desired actions from the combination of objects and impressions. We design experiments to conduct our approach and to see the performance of utilizing object and impression to estimate desired action for images.

The contribution is this paper can be summarized into two points:

(1) We propose the problem of estimating desired actions that are stimulated by annotated images.

(2) To overcome lack of understanding gap between computer and images, we propose to decompose this problem into two sub-problems, by which plenty of external resources and information are available for learning.

By accurately estimating desired actions from images, we could use find images that could stimulate certain desired actions and then use these images for commercial advertisement, presentation, PSAs (Public Service Announcement), image sharing in SNS (social network systems), and so on.

In the rest of the paper, related work is discussed in Section 2.. Section 3. defines the problem investigated in this research. Our approach is presented in Section 4., and followed by experimental explanation in Section 5.. Finally, we draw a conclusion and discuss future work in Section 6..

## 2. Related Work

### 2.1 Visual Analysis

Visual analysis has a long history in the field of visual computing. Regular properties of images includes quality, aesthetic, and objects are widely researched [13] [21]. Recent years, more properties have attracted researchers' attention, such as sentiment [23], interestingness [6], popularity [8], and memorability [9]. These properties involves users as an important factor and computers are trained to simulate as human's thought.

When mentioning images' effect to user behavior, we could naturally come up with advertisement images. However, researches in advertisement related field have few efforts in predicting motivated behaviors from images. Instead, they focus on images' impact on consumers' behaviors [12] and how to improve advertisement quality by matching technology [22].

There are researches in the field of psychology which are trying to find image's influence to user behavior. For example, Patterson [16] have done surveys among Pinterest users and exams how Pinterest images affect users' motivation and behaviors. Their results show that images can do influence user's future behaviors. In this research, we are trying to use technology of visual analysis to find the bridge between images and possible future behaviors that are motivated (desired action as we call).

### 2.2 Deep Learning

Conventional neural network has a long history in fields. In early researches, it has been used for digit recognition with supervised back-propagation networks and successful results were achieved [11]. Recent years, it got a lot of attractions, especially when it is applied on large benchmark datasets consisting of more than one million images. Krizhevsky et al. [10] has shown that deep convolutional neural networks (CNNs) is able to achieve great performance improvement and efficiency for classification on similar datasets such as ImageNet [18].

## 3. Preliminaries

In this section, we will first introduce definition of desired action that we study, and then explain the problem we address in this research.

### 3.1 Desired Action

From viewers' perspective, there is a gradual process when we look at an image, from perception to affect. In perception, there is also a hierarchical model based on Greisdorf's model [5], which includes primitive features (color, shape, texture), objects (person/thing, place/location, activity, event), and inductive interpretation (symbolic value, prototypical displacement). After perception, affect will be stimulated. Affect includes and is not limited to emotion and feeling. Action (or behavior in some researches) is also a product of image viewing. Before action, the feeling of desire to perform the action is belonged to affect and we call

it desired action. Desired action (also referred as DA in this paper) is defined as:

[Definition 1] (Desired Actions) Desired actions are actions that a user wish to carry out after viewing an image.

The reason we want to research on desired action rather than real action is that we want to decrease difference of actions brought by feasibility. For example, action of "eating" is easier to conduct than action "traveling", and it results in the fact that the possibility of performing "eating" is much higher than "traveling" as a result of viewing images. However, the desire of "traveling" could be in similar level with desire of "eating".

### 3.2 Problem Definition

The problem of this research can be defined as: given an image $p \in P$, an image set $S$ with annotation data (including tags, comments, viewer number, etc.), after a series of computing, we could find the desired actions $\{a\}$ that might be stimulated by the image and the degrees of desires:

$$f(p) = \{(a, d) | a \in A, d \in \mathbb{R}^+\}.$$

The image set $S$ is a large set of images that were taken by different users and viewed by others with comments. Usually, these comments include viewers' impression about these images in terms of adjectives.

## 4. Approach

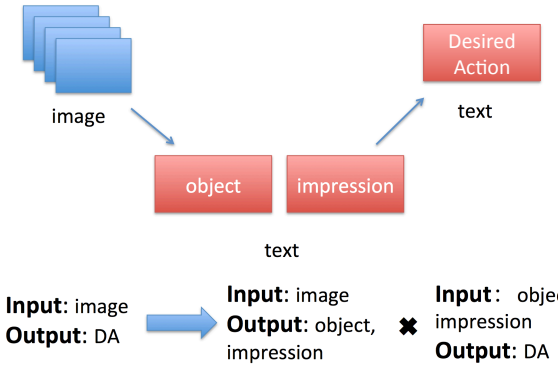


Figure 2 Problem decomposition: match from image to objects and impression, and match from objects and impression to desired action. Objects, impression and desired action are in form of text.

The problem in this research can be mainly summarized as a problem of matching from image to desired action. It can also be treated as an image classification problem if we replace object with desired action in the object recognition problem. Recognition and detection of objects such as "cat" and "sky" has been studied extensively with relatively high performance in the field of computer vision, especially by conventional neural network these years. However, it remains difficult to model impressions indicated by adjectives like "delicious" and "amazing". Chen et al. [3] addressed the big "affective gap" between the low-level visual features and the high-level sentiment when they tried to modeling adjectives correlated with visual sentiments. Similar to sentiment analysis, desired action is more challenging because desired actions correspond to high level abstractions from a given image, which may require viewer's knowledge beyond the image content itself. Meanwhile, a more challenging task in this problem is the lack of data which indicates image and its corresponding desired actions. And the fact that each class (desired action) contains much more diverse images add the difficulty to discover features which can distinguish much more diverse classes from each other.

Imagine we have only 10 images that are corresponding to desired actions "eating", and these ten images contains different types of food. It is quite challenging to find visual features to directly match from images to desired action "eating". What if we could find the common knowledge of those images? From we human's eyes, it is easier to get that those images are all about food, and they all looks very delicious. However can we make computer recognize the same things? The answer is we can utilize knowledge obtained in resources other than these ten images and this knowledge is more confident because of much more data. This is the key idea of our approach, which is to decompose desired action estimation problem into two sub-problems: to extract objects and impressions from images, and to estimate desired actions from the combination of objects and impressions. The reason we choose objects and impressions as a bridge is the fact that they both have implications for actions. Figure 2 also exhibits the main idea.

### 4.1 Extracting Objects and Impressions from Images

Object detection from images has a long history in visual computing literature, and deep learning is considered the state of the art with quite high performance in terms accuracy [21] [19]. In this research, we are going to use existing state of art research to detect objects from images. 1000 classes model in ImageNet [10] is used with regular conventional neural network (CNN) for object detection.

We use deep conventional network to train impression from images. We use social clues attached to annotated images to acquire labeled images. As addressed in research of image sentiment analysis [23], to obtain highly reliable labeled instances is nontrivial, let alone a large number of them. In this research, we utilize users' comments to those images. We first extract adjectives from comments and use those adjectives as viewers' impressions. In total, 888 classes of impressions are extracted with more than 10 images for each class in the training. Then these images are transformed into

format following Caffe's requirement.

We describe briefly the overall architecture of the deep conventional neural networks for training the impression classification, called ImpressionNet. The architecture mostly follows [10]. The while net consists of eight main layers (convolutional and fully-connected) with weights with first five convolutional and other three fully-connected. Image values propagate through the five convolutional layers with pooling, normalization and ReLU. And the three fully-connected layers are used to determine the final neuron activities. Neurons in the fully-connected layers are connected to all neurons in their previous neuron. For each layer, including convolutional layer and fully-connected layer, its output is applied with Rectified Linear Units (ReLUs) [14]: $f(x) = max(0, x)$. Local response normalization layers locate following the first and second pooling layers.

Table 1 The size of input and output data

| Name | Size |
|---|---|
| input | $3 \times 256 \times 256$ |
| data | $3 \times 227 \times 227$ |
| conv1 | $96 \times 55 \times 55$ |
| pool1 | $96 \times 27 \times 27$ |
| norm1 | $96 \times 27 \times 27$ |
| conv2 | $256 \times 27 \times 27$ |
| pool2 | $256 \times 13 \times 13$ |
| norm2 | $256 \times 13 \times 13$ |
| conv3 | $384 \times 13 \times 13$ |
| conv4 | $384 \times 13 \times 13$ |
| conv5 | $256 \times 13 \times 13$ |
| pool5 | $256 \times 6 \times 6$ |
| fc6 | 4096 |
| fc7 | 4096 |
| fc8 | 888 |
| label | 1 |
| output | 1 |

Size of input and output data is shown in Table 1. All images are resized to $256 \times 256$ first and then cropped random $227 \times 227$ patches from the $256 \times 256$ images. The network is then trained on the cropped patches.The output of last fully-connected layer is fed to a 888-way softmax which produces a distribution over the 888 class labels.

Table 2 The shape of each layer

| Name | Shape |
|---|---|
| conv1 | $96 \times 3 \times 11 \times 11$ |
| conv2 | $256 \times 48 \times 11 \times 11$ |
| conv3 | $384 \times 256 \times 3 \times 3$ |
| conv4 | $384 \times 192 \times 3 \times 3$ |
| conv5 | $256 \times 192 \times 3 \times 3$ |
| fc6 | $4096 \times 9216$ |
| fc7 | $4096 \times 4096$ |
| fc8 | $888 \times 4096$ |

Table 2 indicates each layer's shape. The first convolutional layer filters the $227 \times 227 \times 3$ input image with 96 kernels of size $11 \times 11 \times 3$ with a stride of 4 pixels. The output of the first convolutional layer is processed after pooling and response normalization and then taken as input of the second convolutional layer with 256 kernels of size $5 \times 5 \times 48$. The following three convolutional layers are connected to each other without output processing like pooling or normalization. Each of the three fully-connected layers has 4096 neurons.

### 4.2 Estimating Desired Actions from Objects and Impressions

There are many resources that provide a wealth of information to computers about how people use and understand the world in terms of text, such as Wikipedia and Wiktionary. And the prerequisite is that the knowledge can be compiled into a useful representation. In this work, we will utilize ConceptNet [20], which is a project that creates representation of crowd-sourced knowledge. It provides a large semantic graph that describes general human knowledge and how it is expressed in natural language. In ConceptNet, relations are referred as how words are related through common knowledge. Several relations are used especially for this work, as shown in Table 3.

ConceptNet provides a connection to bridge from objects to potential actions. However, in our research, impression is also an important factor that decides desired action, especially degree of the desired action. And ConceptNet lacks of this kind of information. To match from impression to desired action, we will need to add object, since object indicated "what desired action can be stimulated" and impression decides "how much a desired action can be stimulated". There are two types of impressions, general impression which can be used for all objects, like good or bad, and specific impression which is specific for certain objects, like delicious or clean. We are not going to distinguish them in this stage, and our focus will lay on general impression.

Considering impressions, we have to pay attention that different saying of impressions might indicate various degree of desire. For example, we can tell from "Yummy!" and "Yummy!!!!!!!!!" that the later one expresses much stronger of desired action "eating". Another example is "I want to go!" and "I would like to book a flight and fly there now!". Many types of indicators imply the difference of impressions. To make it simple, we will take some explicit forms into account, such as repeated characters in a word ("cooooooool") [2] and usage of special symbols and emojis [1].

Table 3 Relations of objects and actions used in this work

| Relation | Description | Example: (A, B) |
|---|---|---|
| UsedFor | A is used for B; the purpose of A is B. | (ingredient, cook) |
| CapableOf | Something that A can typically do is B. | (cut, knife) |
| MotivatedByGoal | Someone does A because they want result B; A is a step toward accomplishing the goal B. | (arrange, comfortable house) |
| ObstructedBy | A is a goal that can be prevented by B; B is an obstacle in the way of A. | (bad lung, stop smoking) |
| CreatedBy | B is a process that creates A. | (food, cook) |

## 5. Experiment

Corresponding to the approach, the experiment is divided into two parts: extracting objects and impressions from images, and estimating desired actions from combination of objects and impressions.

### 5.1 Object and Impression Extraction

We first extract impressions from annotated images by keeping adjectives in viewers' comments.

We utilize dataset in [17] which contains 1000 categories and 1.2 million images and then apply Caffe [7] for object detection. For impression detection, we use crawled more than 161513 images from the Flickr, each with more than five comments. The number of impressions extracted from those comments is 4301. Then we calculated number of images for each impression class and filtered impressions with less than 10 images. As a result, 888 impressions were kept with more than 153 thousand images in total. Those images were divided into two parts, 80% for training and 20% for testing and validation.

We set the batch size of image as $256 \times 256$ and cropping it into $227 \times 227$ for training. The regression objective is minimized by stochastic gradient descent with a batch size of 256 examples, momentum of 0.9, and weight decay of 0.0005. The learning rate is initialized to 0.01. We run a total of 250,000 iterations.

### 5.2 Match between Objects and Desired Actions

Table 4 Example of relation "CauseDesire"

| Concept | Action |
|---|---|
| bad weather | go somewhere |
| be bore | surf net |
| good weather | travel |
| hatred | punch someone |
| be hungry | cook meal |
| loneliness | make phone call |

In the current stage, we got result from ConceptNet with relation between concepts and actions. Table 4 shows some results about relation CauseDesire. Here, the concepts include both objects and impressions. From this result, we see that the concepts that cause the desire of a certain action is not that proper for the situation of image viewing. The reason is this result comes from people's real feeling in real life, which including a lot of scenarios where image viewing or other visualization is not necessary. This indicates that we still need other approach to improve the performance of matching from objects and impressions to desired actions, in which image viewing is significant.

### 5.3 Result and Discussion

Figure 3 shows some results of impression extraction. Those impressions indicates how users who looked at the image describe about the image.

Table 5 Result: accuracy of impression prediction

| Accuracy | ImpressionNet | Random |
|---|---|---|
| Precision@1 | 0.0148 | 0.0056 |
| Precision@5 | 0.1632 | 0.0281 |

Table 5 shows the accuracy of impression prediction. We use two scores, "Precision@1" and "Precision@5". "Precision@1" means only if the predicted top one result is exactly the same as testing result, the score is 1, otherwise the result is 0. And "Precision@5" means as long as the testing result is among the top five results, the score is 1. We compare our result with randomly give impression. From the result, we can see that the accuracy of impression detection cannot match object detection, but our approach still outperform randomly impression detection. In the experiment, we conducted single-class classification, which will also decrease the accuracy of prediction.

## 6. Conclusion

In this paper, we introduced our proposal of problem: estimating desired actions from annotated images. We demonstrated the difficulty of directly matching from images to desired actions, and thus proposed our idea of decomposing this problem into two sub-problems. Object and impression are selected as a bridge between images and desired actions, and the problem becomes to detect object and impression from images, and estimate desired action from object and

(a) Impressions: delicious, nice, spicy, delightful, thai

(b) Impressions: outstanding, professional, great, vibrant, compelling

(c) Impressions: great, sweet, cute, lovely, adorable

Figure 3 Examples of impressions extracted from viewers' comments annotated with images

impression. Object and impression are in the form of natural language. The advantage of this approach is that we can have much more resources to learn and that common-sense knowledge are introduced to help computer understand images in the perspective of people.

## 7. Acknowledgement

This work was supported in part by Grants-in-Aid for Scientific Research (Nos. 15H01718, and 26700009) from MEXT of Japan.

### References

[1] Becker, L., Erhart, G., Skiba, D., Matula, V.: Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In: Second Joint Conference on Lexical and Computational Semantics (* SEM). vol. 2, pp. 333–340 (2013)

[2] Brody, S., Diakopoulos, N.: Cooooooooooooooollllllllllll-lll!!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In: Proceedings of the conference on empirical methods in natural language processing. pp. 562–570. Association for Computational Linguistics (2011)

[3] Chen, T., Borth, D., Darrell, T., Chang, S.F.: Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. arXiv preprint arXiv:1410.8586 (2014)

[4] Goldstein, E.: Sensation and perception. Cengage Learning (2013)

[5] Greisdorf, H., O'Connor, B.: Modelling what users see when they look at images: a cognitive viewpoint. Journal of Documentation 58(1), 6–29 (2002)

[6] Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., Van Gool, L.: The interestingness of images. In: Computer Vision (ICCV), 2013 IEEE International Conference on. pp. 1633–1640. IEEE (2013)

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

[8] Khosla, A., Das Sarma, A., Hamid, R.: What makes an image popular? In: Proceedings of the 23rd international conference on World wide web. pp. 867–876. ACM (2014)

[9] Khosla, A., Raju, A.S., Torralba, A., Oliva, A.: Understanding and predicting image memorability at a large scale (June 2015)

[10] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)

[11] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation 1(4), 541–551 (1989)

[12] Malik, M.E., Ghafoor, M.M., Iqbal, H.K., Ali, Q., Hunbal, H., Noman, M., Ahmed, B.: Impact of brand image and advertisement on consumer buying behavior. World Applied Sciences Journal 23(1), 117–122 (2013)

[13] Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 2408–2415. IEEE (2012)

[14] Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10). pp. 807–814 (2010)

[15] Ogden, C.K.: The ABC of psychology. Routledge (2013)

[16] Patterson, M.: What happens after the pin? examining how pinterest influences users' motivation and behavior (2015)

[17] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision pp. 1–42 (2014)

[18] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115(3), 211–252 (2015)

[19] Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Networks 61, 85–117 (2015)

[20] Speer, R., Havasi, C.: Conceptnet 5: A large semantic network for relational knowledge. In: The People's Web Meets NLP, pp. 161–176. Springer (2013)

[21] Szegedy, C., Toshev, A., Erhan, D.: Deep neural networks for object detection. In: Advances in Neural Information Processing Systems. pp. 2553–2561 (2013)

[22] Yeo, B.L., Gangjiang, L.: Using image match technology to improve image advertisement quality (Jan 25 2013), uS Patent App. 13/750,552

[23] You, Q., Luo, J., Jin, H., Yang, J.: Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI) (2015)