Interactive Image Annotation with Visual Feedback

Julia Moehrmann and Gunther Heidemann

Institute of Cognitive Science, University of Osnabrueck
{firstname.lastname}@uos.de

Abstract. A semi-automatic process, which support users in the task of annotating large image data sets, has been proposed recently. Images are clustered automatically according to similarity and are presented to the user as a sorted set. During the annotation process, partial annotations are used for further improvement of the clustering. This interactive annotation process has three important properties: First, the user is actively supported in the annotation process. Second, the basis for the clustering is visualized, thereby allowing users to understand what the underlying image features are capable of representing and distinguishing. Third, as the clustering is interactively improved certain categories may turn out to be ineffectual. We will discuss how an interactive annotation system may help to bridge the semantic gap by enhancing the users' understanding of the underlying functionality and how the user and the learning system interact.

1 Interactive Image Annotation

A semi-supervised learning approach for interactive image annotation has recently been presented by the authors in [1]. The semi-automatic user interface for the efficient annotation of image data sets clusters images according to similarity using different image features. Image annotation cannot be automated for the task of creating ground truth data for computer vision systems since correctness is crucial. Therefore the interaction between user and system is of tremendous importance.

The interactive annotation process is displayed in Figure 1. Image data is initially clustered according to similarity using different image features in a Bag-of-Features (BoF) approach. The clustered data is then presented to the user in an optimized user interface (UI) which facilitates the annotation of these images with custom categories. As the user annotates the images, partial annotations are used to learn a better clustering in a semi-supervised process. Identical and different annotations are interpreted as must-link and cannot-link constraints respectively. This allows the calculation of a weight vector w_c for the BoF-feature vectors of category $c \in C$. The category specific vector w_c is calculated to reduce the distances between images of the same category and to increase distances between images of category c and those of category g with $g \in C \land g \neq c$. Images which have not yet been annotated are assigned the weight vector of the

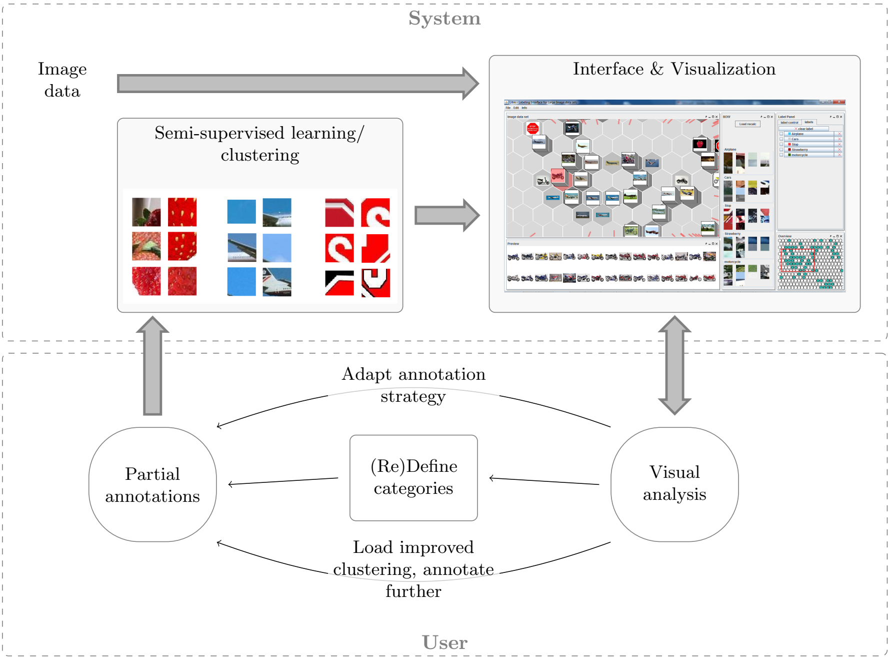
System
Image data
Interface & Visualization
Semi-supervised learning/ clustering
Adapt annotation strategy
Partial annotations
(Re)Define categories
Visual analysis
Load improved clustering, annotate further
User

Fig. 1. Interactive, semi-supervised clustering and annotation process, displaying interaction between automatic components (system) and users, as well as visual word visualization (clustering box).

best matching category. Thus, the ordering of the remaining images is iteratively improved. However, as can be seen in the lower box in Figure 1, which displays the users' actions, alternative events are possible.

1.1 Visual word feedback

A positive side effect of the semi-supervised learning of category-specific weight vectors is that large elements of these vectors $w_c, c \in C$ can easily be identified and set into correlation with the original image features and the regions they were extracted from. This allows us to crop image regions which correspond to high weighted elements in the BoF-vector and use them as a so called *visual word* visualization. We take this even further by using these cropped images to provide the user with information about the decision-making basis of the (semi-supervised) clustering algorithm. While image features in computer vision are in general abstract and difficult to relate to, our visual words allow a simple and clear understanding since they show simple image regions. In the UI those visual words with the highest weights are displayed for each category. This visualization intuitively enhances the users' understanding of computer vision capabilities.

The semantic gap is usually considered as being the missing link between low-level signals (e.g. edges) and high-level concepts (objects). This gap remains a problem despite extensive research in recent years. The major problem in the application of computer vision techniques is that results seldom meet the users' expectations. Especially when we talk about users who are non-experts and have little knowledge in the area of computer vision we have to acknowledge that they do not understand on which basis the classification takes place. While efforts have been made to visualize image features which form the basis for decision-making ([2, 3]), these results are still difficult to interpret by non-experts. Providing visual words which display important parts of an object - as in the semi-supervised learning box in Figure 1 - can however provide intuitive feedback about what the computer is able to extract from the images. In the given examples, the system is capable of extracting important structural information for the categories *strawberry* (left) and *stop sign* (right). A simple effect of understanding would be that the vision system does not read the word "stop" to identify the stop sign but instead uses color or the characteristic edges and corners for the recognition. Another effect of understanding the capabilities of a computer vision system is the visualization of common, i.e., representative image regions which are, however, not discriminative for the actual category. This is the case for the *airplane* category where blue sky regions are displayed.

2 Adaptation of interaction strategies

The visualization of visual words for each category has two additional advantages:

1. Users may adapt their annotation strategy based on this feedback.
2. The classifier performance may be predicted beforehand.

For the first case, let us consider the exemplary visual words given in Figure 1 (semi-supervised learning box). The categories *strawberry* and *stop sign* are well represented by the given examples. The category *airplane* (middle) is mainly represented by blue sky. Based on this feedback users might notice that they annotated only few clusters of this category, which were very similar, i.e., contained blue sky. In order to receive an improved clustering which is capable of distinguishing the categories well, users may adapt their annotation strategy and specifically annotate airplane images with different background. This annotation step will take only very little time but may result in a significantly improved clustering and thereby speed-up the annotation of the remaining images.

The second case, the prediction of classifier performance can also be deduced from the visual words, however, only after a significant amount of the image data was annotated. If, in a learning iteration near the end of the annotation process, the clustering cannot be improved significantly, this may be a hint as to insufficient training data, e.g., if the visual words display blue sky regions even after all images were annotated. The underlying image features do, of course, have an impact on the classification performance. Since the annotation tool uses different image features which represent different image properties this impact should, however, be negligible. In case of insufficient or skewed image data sets, neither the semi-supervised annotation system nor a state-of-the-art classifier has a chance of reaching high recognition rates. In such a case steps can be taken to

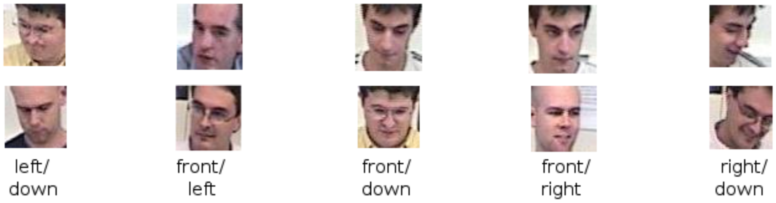

Fig. 2. Difficult head poses for annotation with discrete categories. Faces extracted automatically from PETS video data [4].

improve training image quality before effort is put into classifier selection and calculation.

Another aspect concerning the prediction of classifier performance is the definition of image categories. Let us consider the example of annotating face images according to the head pose. The initial idea is to assign each image to one of four categories: front, right, left, down. If these categories are given, the recommendation by the clustering may aid the user in the decision about the head pose. However, certain head poses are very difficult to decide on and may not fit into one of the given categories, as can be seen in Figure 2. But this is not only a difficult decision for the user. The head poses may simply be too difficult to distinguish, given only four categories. The semi-supervised clustering process helps users understand this by presenting "uncertain" images in-between the two categories they may possibly belong to. For example, after loading a recalculated clustering the images belonging to categorie *left* and *down* would be well separated. The images marked *left/down* in Figure 2 would be displayed between those two categories and thereby intuitively indicate the inability to assign these images to either one category. The annotation system thereby supports the user in adapting the annotation strategy regarding the category definitions. From the clustering and the visual word visualization for each category users may find a tremendous improvement in the category representation when using more than the initial categories, e.g. intermediate categories like *lower-left*. We are aware that head pose annotation is a very specialized problem since it tries to assign continuous data to discrete categories. However, this annotation is a real world problem which has to be performed. The problem of possibly inappropriate categories also arises with other recognition tasks where no clear separation is available or where the variance in object appearance is larger than expected.

3 Conclusion

The semi-supervised image annotation system exploits user interactions with respect to the sorting of the data. Additionally, visual word representations are provided to the user as a basis for further interactions with the underlying system. These interactions may have an impact on the training data and the image categories. The interactive process might already yield information about classifier performance during ground truth annotation and allow for improvements at an early stage in the development process.

References

1. Moehrmann, J., Heidemann, G.: Semi-Automatic Image Annotation. In: Computer Analysis of Images and Patterns. Lecture Notes in Computer Science, vol. 8048, pp. 266–273 (2013)
2. Shoukry, L., Klenk, S., Heidemann, G.: MPEG-7 Feature Visualization for CBIR Systems. In: Proc. Int. Conf. on Computer Theory and Applications (ICCTA 2010). pp. 86–90 (2010)
3. Vondrick, C., Khosla, A., Malisiewicz, T., Torralba, A.: HOGgles: Visualizing Object Detection Features. ICCV (2013)
4. Ferryman (ed.), J.: Fourth IEEE Intl. Workshop on Performance of Tracking and Surveillance (PETS-ICVS). Graz, Austria. http://www.cvg.rdg.ac.uk/PETS-ICVS