Annotating images with suggestions — user study of a tagging system

Michal Hradiš, Martin Kolář, Aleš Láník, Jiří Král, Pavel Zemčík and Pavel Smrž

Faculty of Information Technology
VUT — Brno University of Technology
Brno Czech Republic

Abstract. This paper explores the concept of image-wise tagging. It introduces a web-based user interface for image annotation, and a novel method for modeling dependencies of tags using Restricted Boltzmann Machines which is able to suggest probable tags for an image based on previously assigned tags. According to our user study, our tag suggestion methods improve both user experience and annotation speed. Our results demonstrate that large datasets with semantic labels (such as in TRECVID Semantic Indexing) can be annotated much more efficiently with the proposed approach than with current class-domain-wise methods, and produce higher quality data.

Keywords: Restricted Boltzmann Machine, human-assisted learning, user interface, image tagging, crowdsourcing, image classification

1 Introduction

[1] Automatic or semi-automatic image tagging, classification, and semantic analysis is one of the important and open problems of the contemporary image data management.

Obtaining high-quality annotations for large image and video datasets is paramount in the area of all types of classification and especially semantic classification [2]. The annotated datasets are being used not only to learn the semantic classifiers, but they are needed also for evaluation of different approaches, and for comparisons of the results in order to reliably identify and evaluate the promising methods. We propose the idea that such datasets including large number of semantic categories could be efficiently obtained through annotation of one image or video shot at a time (*image-wise* tagging) provided the system suggests

[1] Paper will be presented at ACIVS 2012 and published in LNCS. The original publication is available at www.springerlink.com.

[2] Semantic classification of images is understood as an assignment of semantic tags to images or parts of images. The tags can represent objects (e.g. car, person, building), conditions (e.g. sunny, winter, outdoor, fog), activities (e.g. singing, dancing, running), or possibly other relevant semantic categories.

the likely tags based on the content, tags assigned to near-by images, and also based on tags already assigned to the currently annotated image or video.

ITS (*Intelligent Tagging System*), the web-based image tagging system we implemented for this purpose, suggests tags for an image or video by modeling dependencies between tags assigned to an image using Restricted Boltzmann Machines [3] (RBM), and through utilizing tags of temporally collocated images from the same gallery. The objective of the tag suggestion methods is to allow *image-wise* tagging (assign tags to an image) rather than *class-domain-wise* tagging (assign images to a tag). According to the user testing we performed, this approach makes tagging faster, easier to use, more intuitive, and more precise. The produced dataset contains significantly more annotations of infrequent tags, since it makes tagging of rare classes more probable compared to the class-domain-wise tagging.

Several existing datasets contain enough tagged images to make learning and comparison of semantic image classifiers possible [11, 4, 12, 6, 5]. These datasets were annotated using various tagging methods, typically by creating a taxonomy and adding positive and negative examples in each class manually, or by searching on Internet and checking the search results by hand [4, 12]. Each image in such datasets is assigned only to a single class, which inhibits class correlation analysis. When attempting to find tag correlations, for example for the suggestion of co-occurring tags, data generated this way cannot be used.

Alternatively, TRECVID semantic indexing dataset [9] is annotated by *Active Learning* [1] and contains annotations of possibly 500 tags for each video-shot; however, positive examples are rather sparse in this dataset. The Active Learning is a class-domain-wise annotation approach. It utilizes a network of classifiers, which are organized in such a way as to take into account a variety of low level features and descriptors. These include text, local and global visual information, as well as conceptual context. The classifiers are iteratively trained on currently available annotations and provide users with examples which would be most informative when annotated for a given tag.

Outside computer vision research, usage of visual media databases becomes more common and the amount of available content grows rapidly. Semantic information in the form of tags greatly improves the ability to search and browse such databases. As opposed to visual information, semantic information is more useful for navigation in the databases; however, it is also much harder to extract from the contents. At the present, reliable extraction of general tag-level semantic information from images is not possible, and state-of-the-art methods provide only mediocre results [9]. Reliable and broad semantic information has to be currently provided by users.

Existing media databases (e.g. Flickr and YouTube) allow users to tag the content they upload by typing words or by selecting from a list of tags automatically suggested based on previously added tags. Methods used in our image tagging system are directly applicable in such databases, and the obtained experimental results are relevant, to an extent, for such applications as well. Among others, the experiments show that tagging using the tag suggestion provides

richer information, and that the users find it more pleasant and straightforward, indicating that users would be more inclined to tag content with good suggestions.

The paper first presents the description of the proposed semi-supervised prediction method and the technical description of the web user interface. The experiments and results are presented in Section 3 together with discussion of the results. Finaly, the paper is concluded in Section 4.

2 Suggestion engine

A key part of ITS is the suggestion engine[3], which makes a prediction of likely tags, given current positive and negative tags on an image. We have combined a method with a global prior on tag co-occurrence (Restricted Boltzmann Machine), with a method for using information from tags in concurrent images (local tag suggestion). We have chosen these methods to make it possible to use the annotation system for various types of data (independent images, related images, video sequences), and in various ways to allow flexibility.

2.1 Restricted Boltzmann Machine

RBM is an undirected bipartite graphical model [3]. It defines a probability distribution over a vector of visible variables $\mathbf{v}$ and a vector of hidden variables $\mathbf{h}$. In the RBM model, the visible variables are independent of each other when the hidden variables are observed and vice versa.

For the purpose of modeling dependencies among semantic tags, the visible variables $\mathbf{v}$, each corresponding to presence of a tag, are binary. In our work, the hidden variables $\mathbf{h}$ are binary as well.

The joint probability over $\mathbf{v}$ and $\mathbf{h}$ is defined as

$$p(\mathbf{v}, \mathbf{h}) = \frac{exp(-E(\mathbf{v}, \mathbf{h}))}{Z}, \tag{1}$$

where Z is a normalization constant and E is energy function given by

$$E(\mathbf{v}, \mathbf{h}) = -\mathbf{v}^\top \mathbf{W}\mathbf{h} - \mathbf{v}^\top \mathbf{b}^v - \mathbf{h}^\top \mathbf{b}^h, \tag{2}$$

where (W) is a matrix of weights between elements of $\mathbf{v}$ and $\mathbf{h}$, and $\mathbf{b}^v$ and $\mathbf{b}^h$ are biases of visible respective hidden variables. Conditional dependencies between the visible and hidden variables are expressed as

$$p(\mathbf{h}|\mathbf{v}) = \sigma\left(\mathbf{W}\mathbf{v} - \mathbf{b}^v\right) \text{ and } p(\mathbf{v}|\mathbf{h}) = \sigma\left(\mathbf{W}^\top \mathbf{h} - \mathbf{b}^h\right), \tag{3}$$

where $\sigma()$ is a sigmoid function.

As a generative model, RBM could be trained using maximum likelihood. However, derivatives of the likelihood are intractable. To overcome this problem,

[3] The ITS system is available at http://medusa.fit.vutbr.cz:15161.

Hinton [7] introduced a practical approximation called *Contrastive Divergence* (CD). The CD algorithm computes gradients for optimization as

$$\nabla \mathbf{W} = \langle \mathbf{vh} \rangle_{data} - \langle \mathbf{vh} \rangle_{recon} \tag{4}$$

$$\nabla \mathbf{b}^v = \langle \mathbf{v} \rangle_{data} - \langle \mathbf{v} \rangle_{recon} \tag{5}$$

$$\nabla \mathbf{b}^h = \langle \mathbf{h} \rangle_{data} - \langle \mathbf{h} \rangle_{recon}, \tag{6}$$

where $\langle . \rangle_{data}$ are expectations with respect to the distribution of data and $\langle . \rangle_{recon}$ are expectations with respect to the distribution of reconstructed data. The reconstructed data is obtained by starting with a data vector on visible variables, and sampling first from distribution $p(\mathbf{h}|\mathbf{v})$ and then $p(\mathbf{v}|\mathbf{h})$ (Equation 3).

In the context of tag suggestion, the task of RBM is to provide marginal probabilities of unobserved tags which constitute the visible variables $\mathbf{v}$ as more and more tags become observed (by actions of a user). Several algorithms could solve inference in in the RBM model. We chose *Gibbs sampling* which draws several samples from the RBM distribution. The means of marginal distributions $E(p(v_i))$ can then be computed from the the samples. Gibbs sampling starts by assigning random values to unobserved variables, and a sample is obtained by iterating between computing $p(\mathbf{h}|\mathbf{v})$ (Equation 3) and sampling from it, followed by computing $p(\mathbf{v}|\mathbf{h})$.

As it is not practical and/or desirable to obtain a large training dataset where presence or absence of all tags for all images would be known due to a large number of possible tags (hundreds or thousands), inevitably, such dataset has to have sparse annotations, and the learning algorithm has to handle situations where potentially large portion of the tags is unobserved. Several methods for handling missing training data in the context of RBM were proposed. Single missing value can be easily filled by sampling from its exact conditional distribution (it is known for single unobserved variable). More missing values can be treated in the same way as the other parameters [8] if they are updated often during learning. This approach is efficient only on training sets of limited size. Salakhundinov et al. [10] introduced a radical way of dealing with missing values by using RBMs with different numbers of visible units for different training cases. This approach is able to handle very sparse data; however, it no longer produces a single RBM model.

In our work, we decided to use Gibbs sampling to fill the unobserved values in the training data. For the CD gradients (Equation 4), the data means $\langle . \rangle_{data}$ have to be computed. This can be done by drawing samples from the distribution of the unobserved visible variables conditioned on the observed visible variables. This distribution is not known during learning of the RBM model. However, current imperfect RBM model can be used instead as an approximation. When a sample from the distribution of the visible data is obtained, the CD algorithm proceeds exactly as described in Section 2.1.

2.2 Local tag suggestion

Aside from the RBM suggestion method, tags are also suggested if they are positively annotated in nearby images in the gallery. A gallery is viewed as a

chronological sequence, with images $\{I_i\}_{i=1}^{N}$. When generating suggestions for a given image I_i, each tag is given a weight ω, given by

$$\omega = \sum_{i=1}^{N} \frac{1}{log(|p-i|+1)} * has\_tag(I_i), \tag{7}$$

where

$$has\_tag(I_i) = \begin{cases} 1 & \text{if the tag is positively annotated on } I_i \\ -1 & \text{if the tag is negatively annotated on } I_i \\ 0 & \text{if the tag is not annotated on } I_i \end{cases}$$

The $\frac{1}{log(|p-i|+1)}$ term ensures that closer annotations have more weight on ω, and the $has\_tag(I_i)$ term ensures that positive annotations have positive weight, negative annotations negative weight, and all others are ignored. Tags are then ordered by their ω from highest to lowest. Any tags with $\omega > 0$ are then suggested, in this order.

2.3 Integration of Suggestion and User Interface

When suggesting n tags, $\lfloor n/2 \rfloor$ are from the RBM model, $\lfloor n/2 \rfloor$ from local tag suggestion, and if n is odd the remaining one is chosen with either method with equal probability. That ensures that when only one tag is being added, neither method is favoured.

When an image is loaded, 15 tags are chosen and three annotating options are available to the user. As seen in Figure (picture of the web), they are as follows:

1. Each of the 15 suggested tags is presented with a "check" and a "cross". When clicking check, the tag is added as positive annotation, the cross adds negative annotation. When clicking either, the tag dissapears from the suggestion list, and a new one is added at the end of the list.
2. The user can use an autocompleting text field, where any typed word or part of a word is matched with all occurrences in existing tags as a substring. For example, when typing person, the user is presented with "person", "male person", "female person", and others. This ensures that when no information is given yet, the user can easily add information that's compatible with the current collection of tags in the database. When any of these is clicked, it gets added to the current suggestion, and the suggested tags are refreshed accordingly. Users are allowed to enter new tags which are not yet in the database; however, such tags are not immediately considered by the RBM model. It is more appropriate to add new tags to the RBM model when the number of positive annotations of such tags increases over certain threshold in order to prevent saturating the model by rare or otherwise irrelevant tags.
3. Given the chronological sequence of images, three preceding and three succeeding images are shown on the right. When any of these is clicked, the positive tags that have been annotated on that image are copied over to the current image, and the suggested tags are refreshed accordingly.

Fig. 1. Typical view of the ITS web interface. Annotation option parts are outlined in red.

The suggestion operation takes on average 0.1 seconds, making the system responsive and allowing quick interaction with the user. In case of sequential video frames, this interface allows users to seamlessly copy tags from previous images to the current one, either by copying tags from the three preceding and three succeeding images, or by selecting the suggested local tags. Another use scenario is the annotation of holiday photos with recurring themes, people, and elements. In the case of unusual images and tags that are not a priory likely, the RBM suggestions may not be accurate very useful at first; however, by providing one or several tags relevant to the image (e.g. by using the autocompleting text field) will make co-occuring tags likely to be suggested.

3 Experiments and Results

In order to identify the usability and usefulness of our system, we performed two experiments with users: testing with untrained individuals with minimal

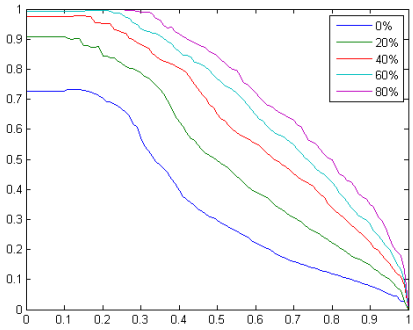

Fig. 2. Precision-Recall Curves of tag suggestion for different numbers of known tags per image. The curves are for different probabilities that the tags in the TRECVID 2011 semantic indexing dataset are known.

support, and testing with expert annotators for an extended period of time. In order to make the test replicable, we used only images and tags[4] from the TRECVID 2011 Semantic Indexing task[5], and disabled the feature to add new tags.

Besides the reproducibility of the experiments by others, there are several other advantages of using the TRECVID data. A part of the data is already annotated and can be used to learn the RBM tag-dependency model. Further, the dataset was annotated by Active Learning [2] which provides a baseline for comparison.

In addition to the user study, the ability of RBM to model dependencies among tags and the ability to estimate marginal tag probabilities by Gibbs sampling was tested on the TRECVID data. This experiment gives an objective information of the RBM suggestion system alone.

3.1 RBM suggestion

The RBM tag suggestion was tested on a training dataset for semantic indexing task from TRECVID 2011 evaluations. The dataset consists of 400 hours of video from which over 260 thousand images (key-frames) were extracted. For the dataset, 345 semantic classes were annotated by Active Learning[6] [2]. Total 14M shots-level annotations were collected (approximately 16 %). Note, that only 400 thousand of the annotations are positive. On average, there is over

[4] Examples of the classes are Actor, Airplane Flying, Bicycling, Canoe, Doorway, Ground Vehicles, Stadium, Tennis, Armed Person, Door Opening, George Bush, Military Buildings, Researcher, Synthetic Images, Underwater and Violent Action.

[5] `http://www-nlpir.nist.gov/projects/tv2011/tv2011.html`

[6] `http://mrim.imag.fr/tvca/`

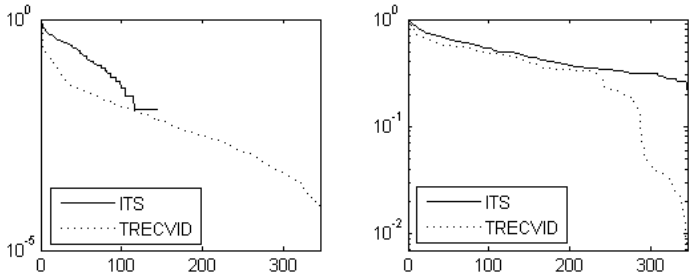

Fig. 3. Normalized numbers of positive (left), respective negative (right) annotations for classes in TRECVID 2011 semantic indexing dataset as annotated by Active Learning [1] and ITS expert users. Scales of y-axes are logarithmic.

1100 positive and 42 thousand negative annotations for each class. Distribution of annotations is shown in Figure 3.

The TRECVID dataset was divided into two parts. First[7] 200 thousand key-frames were used for training. From the remaining 60 thousand key-frames 20 thousand were randomly selected for testing. Key-frames from a single video were assigned exlusively to only one of the sets.

Precision-recall curves (PRC) were computed on the testing part of the dataset for for probabilities 0 %, 20 %, 40 %, 60 % and 80 % that the annotated tags are known - known tags were sampled randomly and independently for each key-frame. Note that even for the high probabilities, only a small number of tags per a key-frame are actually known due to the fact that only 53 tag annotations are available for a key-frame on average.

The optimal strength of L2 regularization (*weight decay*) and size of the hidden layer was selected by grid search using cross-validation. The CD training process iterated 20 times over the training set before terminating, and the marginal probabilities of tags on the testing set were estimated using fifty samples. The actual size of the optimal hidden layer was 64. Adding more hidden variables did not improve modeling tag dependencies, and it reduced the ability of RBM to model a priori probabilities (when no tags are known for an image).

Results in Figure 2 clearly show that the RBM combined with Gibbs sampling can utilize the information provided by the known tags, and that precision significantly improves with the number of known tags. The PRC for 0 % of known tags corresponds to a priori probabilities of tags, and it exhibits the relatively good results due to unbalanced counts of positive annotations of the individual classes.

[7] Videos were sorted according to their titles.

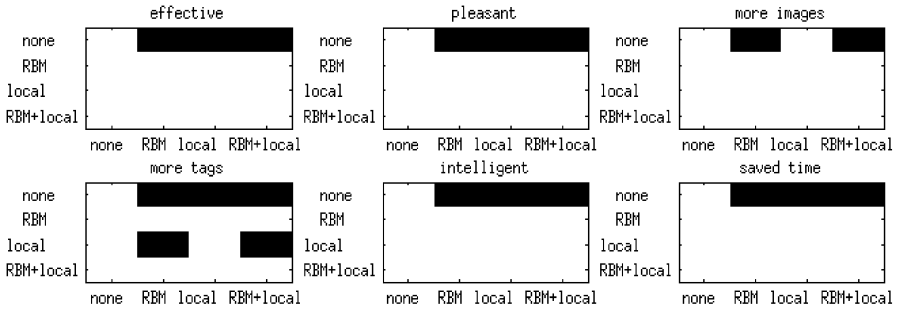

Fig. 4. Black squares represent a significantly better outcome in the user evaluation, according to the questionnaire. The questions allowed a $1-5$ rating on effectiveness, pleasantness, amount of images, amount of tags per image, perceived method intelligence, and whether the method saved time.

3.2 Testing by Untrained Users

10 randomly selected technical university students were asked to use 4 different tag suggestion methods using our system, with as little training as possible. The 4 methods are:

1. **none** — no suggestion method
2. **RBM** — only Restricted Boltzmann Machine suggestion (Section 2.1)
3. **local** — only local tag suggestion (Section 2.2)
4. **RBM+local** — the combination of Restricted Bolzmann Machine and local tag suggestion, as presented in section 2.3

The methods were ordered randomly and the user was not told which is which. After using each method, the user was asked to answer a questionnaire with questions regarding the rating and usability of the method, and data regarding the amount of annotations created was stored.

According to the results (Figure 5), **RBM** and **RBM+local** suggestion methods allow significantly[8] faster annotation. There were no significant differences between **RBM** and **RBM+local**, nor between **none** and **local**. According to the questionnaire, method **none** is found by the users to be significantly[9] inferior to all the other methods in almost all aspects. No other significant differences were found, except that **RBM** and **RBM+local** received better marks in the ability to facilitate annotating more tags per image compared to **local**.

3.3 Testing by Expert Users

Three expert users were asked to use the combined tag suggestion method (Section 2.3). The users previously took part in TRECVID 2011 collaborative annotations [1], and had at least two hours experience with ITS. The users spent

[8] Using the paired t-test at the 10% significance level.
[9] Using the Mann-Whitney U test at the 10% significance level.

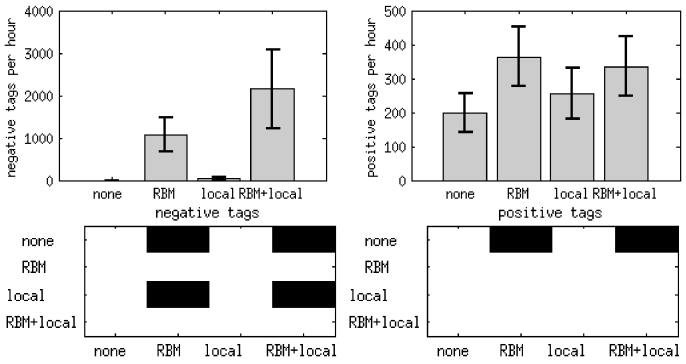

Fig. 5. The top graphs show the mean number of tags assigned per hour with confidence intervals at 90% significance level. The bottom graphs show black squares where the column methods annotate significantly more tags per hour than the row methods.

a total of three hours annotating randomly selected videos from the TRECVID dataset.

In this setting, the number of positive and negative annotations assigned per hour was 448 and 3085 respectively, averaging 13.1 positive annotations per image. The annotating speed compares very favorably to class-domain-wise annotation for which the authors of [1] expect 2 seconds per annotation; moreover, only 2.5% of the annotations in the TRECVID 2011 SIN [9] dataset are positive. Distribution of the annotated tags is shown in Figure 3. When compared to the original distribution of tags obtained by the Active Learning method [1], the ITS tags have a heavier tail distribution for both positive (kurtosis 8.35 in TRECVID and 4.18 by ITS), and negative annotations (kurtosis 2.18 in TRECVID and 1.98 by ITS).

3.4 Discussion

According to the distribution of tags obtained by ITS (Figure 3), infrequent classes are more likely to get tagged with this method than with class-domain-wise Active Learning [1]. One of the probable causes is that users are able to assign the most relevant tags to images using the autocompleting text field even though the tags are not suggested. This is a clear advantage, as positive examples of less frequent classes are hard to obtain by Active Learning, which forces users to asses a huge number of almost random images for each of the infrequent classes. This effect would be even more pronounced if the set of annotated tags was larger.

Another problem of the class-domain-wise Active Learning is that the underlying classifiers may drift according to early examples to a specific type of images which are not representative of a whole class. For example, consider the first annotated images for a dog class happen to be grayscale. The classifier could

focus on the color in such case, and it may never recover. The **local** suggestion method does not exhibit this type of issues. The RBM model could learn inaccurate a priori probabilities on early examples. However, these inaccurate a priori probabilities will get corrected: as tags become more likely, they are more likely to be suggested, and consequently they will become more likely annotated as negative.

In our experiments, previously annotated data is used to overcome the cold start problem. If such data was not available, the RBM model could be initialized according to a text corpus. However, it would be feasible to start without any tag dependency knowledge at all, as the **local** suggestion method already allows good suggestions in many situations. Only the speed of annotation would be negatively affected in such case. We suspect that the TRECVID Semantic Indexing (SIN) 2011 dataset does not allow the RBM model to provide as good suggestions as could be reached due to very sparse positive annotations, and we expect that speed of annotation would increase if more densely annotated dataset was used.

In TRECVID SIN, the annotated objects are short video shots. In the class-domain-wise Active Learning [1], a single shot is assessed multiple times by different users for different tags. To make the annotations highly reliable, the assessors would have to view a video-shot again for each annotated tag. Viewing each shot would be very time-consuming. On the other hand, extending ITS to video-shots would introduce only minor overhead, as a shot has to be viewed only once to annotate all tags.

4 Conclusion

We created a system for human-assisted image-wise annotation with tag suggestions which could be used to obtain large semantically labeled datasets. The suggestion methods, as well as the annotating system itself, could be applied in the context of public media databases.

According to the experiments, the proposed method for modeling dependencies of tags using RBM is able to utilize previously assigned tags and estimate marginal probabilities of other tags. Both suggestion methods improve user experience when annotating images, and the RBM model and its combination with local tag suggestion improve annotation speed as well. Experienced users are able to produce positive annotations at a much faster rate using ITS compared to class-domain-wise Active Learning [1]. In addition, the obtained annotations contain a higher percentage of positive examples of infrequent classes.

As a future work we intend to combine ITS with research in image feature extraction and semantic image classification [9]. The RBM suggestion method can be extended to integrate information from the assigned tags with visual information, so that suggestions are made more reliable, especially when no tags are yet assigned. A natural way to combine the information is, for instance, provided by Conditional-RBM models [8]. Further, the current local tag suggestion method could be given a stronger foundation by being integrated in the probabilistic suggestion model as well.

Acknowledgements

This work has been supported by EU-7FP-IST - GLOCAL - EEU - 248984, and BUT FIT grant No. FIT-11-S-2, and Research and Development Council of the Czech Republic - CEZ MMT, MSM0021630528.

References

1. S Ayache. Video corpus annotation using active learning. *Proceedings of the IR research, 30th European*, 2008.
2. Stéphane Ayache and Georges Quénot. Evaluation of active learning strategies for video indexing. *Signal Processing: Image Communication*, 22(7-8):692–704, August 2007.
3. M A Carreira-Perpinan and G E Hinton. On Contrastive Divergence Learning. In Robert G Cowell and Zoubin Ghahramani, editors, *Artificial Intelligence and Statistics*, page 17. Citeseer, Society for Artificial Intelligence and Statistics, 2005.
4. Jia Deng, Wei Dong, Richard Socher, Li-jia Li, Kai Li, and Li Fei-fei. ImageNet : A Large-Scale Hierarchical Image Database. pages 2–9.
5. Mark Everingham, Luc Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, September 2009.
6. G Griffin and A Holub. Caltech-256 object category dataset. 2007.
7. Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.
8. Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.
9. Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Wessel Kraaij, Alan F. Smeaton, and Georges Quéenot. TRECVID 2011 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2011*. NIST, USA, 2011.
10. Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted Boltzmann machines for collaborative filtering. *Proceedings of the 24th International Conference on Machine Learning (2007)*, pp:791–798, 2007.
11. John R Smith, Milind Naphade, Jelena Tesic, Shih-fu Chang, and Winston Hsu. Standards Large-Scale Concept Ontology for Multimedia. *Evaluation*, pages 86–91, 2006.
12. Antonio Torralba, Rob Fergus, and William T Freeman. 80 Million Tiny Images: a Large Data Set for Nonparametric Object and Scene Recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–70, November 2008.