LDADEEP+: LATENT ASPECT DISCOVERY WITH DEEP REPRESENTATIONS

Chieh-En Tsai *Hui-Lan Hsieh* *Winston Hsu*

National Taiwan University, Taipei, Taiwan

ABSTRACT

Nowadays, with the success and fast growth of social media communities and mobile devices, people are encouraged to share their multimedia data online. Analyzing and summarizing data into useful information thus becomes increasingly important. For online photo sharing services like Flickr, when users are uploading a batch of daily photos at a time, the tags users provided tend to be rather vague, containing only a small amount of information. For better photo application and understanding, we attempt to automatically discover semantic-rich (hidden) aspects of photos merely by looking at image contents. In this paper, we propose an effective model, which is a combination of LDA model and deep learning representations, to realize the idea of automatic aspect discovery. We then discuss the properties of this aspect discovery model through experiments on event summarization task . In those experiments, we show the high diversity and high quality of aspects discovered by our proposed method. Meanwhile, we conduct an user study to evaluate the quality of the summarized results. Moreover, the proposed method can be further extended to human attribute discovery for a given event. We automatically discover different aspects on our Olympic Games data (e.g. football, ice skating).

Index Terms— Aspect discovery, event summarization, LDA, deep representation

1. INTRODUCTION

With the growth of social media, more and more people like to share their daily life and activities through the Internet. Every minute, thousands of photos and videos are uploaded to photo sharing websites. However, this valuable information resource is cluttered and therefore unusable unless we can find a way to uncover the structures and aspects behind those photos. Event summarization is one way to achieve this goal. However, in a real-world situation, tags may not be precise especially when users are uploading a batch of photos at a time. In a complementary manner, our work focuses on content-based event summarization task: given the result images of searching by tags (without ranking) in a large daily photo dataset, we attempt to automatically summarize all aspects of the event (Fig. 1).

Latent Dirichlet Allocation (LDA) [1] is a powerful, unsupervised, generative probabilistic model. Arora et al. [2] use LDA model as a generative approach to solve multi-document summarization problem. This gave us the inspiration of solving event summarization problem with LDA model in the early stage. The main difficulty of using LDA model in computer vision is to define what stands for a word (visual-word) in an image. If visual-words with deficient semantic meanings are defined, it directly leads to poor performance. Typical solution [3, 4, 5] is to create a codebook of local descriptors to represent *visual-words* and view each image in a bag-of-visual-word fashion.

This is work is supported by grants MOST 104-2622-8-002-002, MOST 103-2911-I-002-001, NTU-ICRP-104R7501, MediaTek, and Intel Corp.

In this paper, we propose a model to deal with this problem, it is based on LDA topic model and takes the advantage of high quality deep learning representations to make the learned aspects more meaningful. When representing an image, instead of building a bag-of-visual-word codebook as standard methods do, we argue that direct utilization of semantically-rich global representations [6] yields a better result.

We conduct an user study and a series of experiments to show that LDADeep+ possesses:

1. the ability to discover high-semantic-level aspects
2. robustness against searching results that are weakly related to the target event
3. high discriminative power on learned aspects
4. the ability to avoid repeated summarization of one aspect

Finally, we show one application (Human attribute analysis) made possible by LDADeep+ to illustrate the importance of good aspect discovery and its benefits. Niebles et al. [7] use Spatial-Temporal words along with LDA model to perform human activity recognition which takes video sequence as input. In Section 3.3, we discover aspects of human attributes, like pose and activity, automatically from non-sequential images. Our contributions include:

- We perform aspect discovery automatically on image contents without the needs of human annotated tags or labels
- We revisit LDA model and arm it with a state-of-the-art deep learning technique to break new ground
- We conduct a user study and examine the high quality aspects it learned
- We apply the proposed model in human attribute analysis to

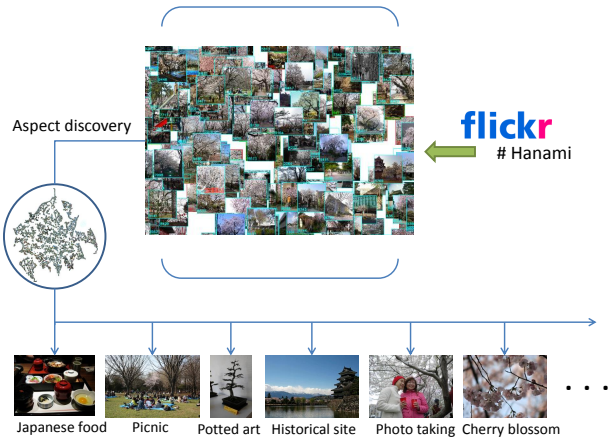


Fig. 1. Given a target event name (e.g, Hanami), we attempt to automatically summarize and discover all aspects (e.g. cherry blossoms, picnic) of the given event from a large collection of daily photos.

2. METHOD

For unsupervised event summarization task, we focus on summarizing all the aspects of an event. We define an aspect as a commonly seen component among images which belong to the same event (e.g. *cherry blossom* and *picnic* are both aspects of Hanami). With this definition, summarizing an event by image content is now equivalent to finding components which are most commonly seen among image data, meanwhile, maximizing the diversity of found aspects.

As discussed in Section 2, the lack of representative power of visual-word is the main difficulty of using LDA model. Nevertheless, with the huge success of deep learning achieved recently [8, 9], revisiting of modeling images with LDA model is worth a try. We propose a model called LDADeep+ which directly utilizes global feature generated by deep neural network: each output neuron represents a semantic-rich visual-word and the normalized output responses are viewed as the visual-word distribution.

The reason we choose not to build a codebook is that deep learning feature is a sophisticated global representation of image, and building a codebook on top of it will lose valuable information for later aspect discovery process. An image **i** can be viewed as a document composed of *N* words picked from a vocabulary set of size *V*, where *V* is the dimension of deep neural network feature. Suppose there are *M* images in a search result *D*, the model regards the generative process for each image **i** in *D* as a simulation of document generative process [1] as follows:

1. Choose $N \sim$ Poisson(ξ)
2. Choose $\theta \sim$ Dirichlet(α)
3. For each of the *N* visual-word i_n:
   (a) Choose an aspect $a_n \sim$ Multinomial(θ)
   (b) Choose a visual-word i_n from $p(i_n|a_n, \beta)$, a multinomial probability conditioned on the aspect a_n

Where α and β are hyperparameters of Dirichlet distributions for aspect distribution per image and visual-word distribution per aspect respectively. The model is suitable for automatic aspect discovery task. The aspects learned will be the most commonly seen combination of visual-words and diversity is guaranteed among these aspects.

3. EXPERIMENTS

We conduct our experiments on YFCC100M [10], a dataset with 100 million of daily photos, videos and their hashtags, if any, provided by Yahoo! Flickr. In order to build a more general model, we represent images by the feature extracted from the second fully-connected layer in AlexNet [9] (fc7 feature, pre-trained on ImageNet [11] dataset), which, as shown in [12, 6], is a feature with natural clustering effect and tends not to overfit to any task. For evaluation, we compare our method with LDA-SIFT, a standard technique using LDA model with SIFT local descriptor and codebook building, and k-means clustering (on fc7 feature). For LDADeep+ and LDA-SIFT model, we render each aspect by the image which has the max response to the corresponding dimension in the latent space; for k-means method, we simply choose the one closest to a centroid.

To test the performance of those models, we conducted a questionnaire with 4 events (Hanami, Holi, Batkid, Occupy Movement), 3 models and 12 summarization results in total for subjects to score on a scale from 1 to 6 on the criteria that the best summarization should *(1) provide as many aspects as possible* and *(2) obtain higher quality for each aspect*, of course without knowing which summarization is generated by which model. In the end, we collected 66 responses. The proportions after scores are turned into ranking are reported in Table 1.

Table 1. Result of the questionnaire. LDADeep+ consistently outperforms other methods except for Holi case.

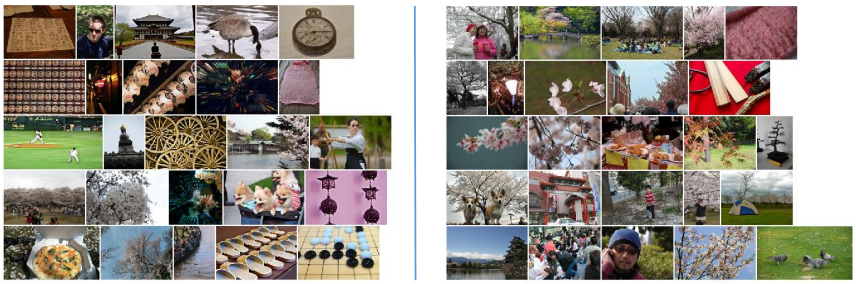
event	LDA-SIFT	k-means	LDADeep+
Hanami	26.3%	22.2%	**51.5%**
Holi	15.0%	**55.4%**	29.7%
Batkid	38.5%	22.1%	**39.4%**
Occupy Movement	31.2%	21.6%	**47.2%**

The numbers shows that people generally favour the summarization generated by the proposed model. In the following subsections, we will delve into the result of user study and elaborate properties of LDADeep+ model. We first conduct experiments to compare different models and different representation of images and finally, briefly conclude our experiments with an application: human attribute analysis, to demonstrate one among many potential applications of our model.

3.1. Comparison of different summarization models

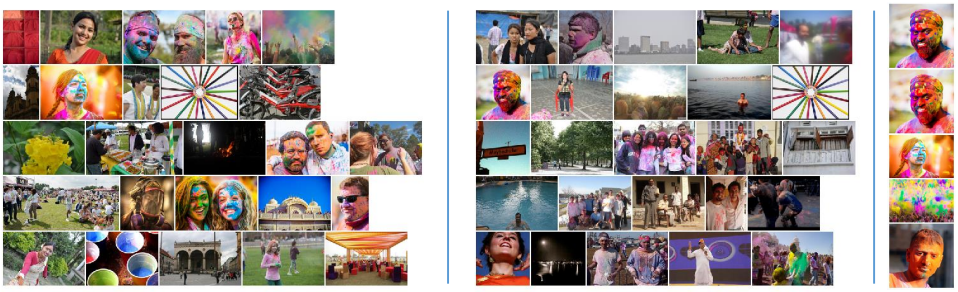
As Table 1 shows, LDA-like model generally outperforms k-means method especially for Hanami event. The main difference between k-means and LDA is that k-means only focuses on finding *the most diverse centroids (aspects)* while the latter also considers *how frequently those aspects appear*. The effect of this difference can be easily observed. In Fig. 2, LDADeep+ discovers lots of aspects strongly related to *cherry blossom*, an essential aspect of Hanami which is commonly seen, while k-means method discover diverse images but most of them are not as strongly related to Hanami as those summarized by LDADeep+. The reason why this effect is significantly shown in Hanami event is that Hanami is a popular event often talked about by people in Japan. This make the search result more likely to contain a large portion of weakly related images and the most diverse centroids of images often lie in the weakly related part. This shows that our method is more robust. When the searching results do not guarantee high relevance, LDADeep+ can still find essential aspects by looking for patterns most commonly appear.

Another strength of LDADeep+ is that it can recognize aspects it discovered and avoid repeated summarizations in advance. It is interesting to find that LDADeep+ outperforms k-means method in all cases except Holi. A crucial aspect of Holi is *colorful* and LDADeep+ summarizes *colorful* by the leftmost image in the second row of Fig. 3(b). Once the *colorful* aspect has been recognized by LDADeep+, images with high response to *colorful* will cluster into one peak in latent space. Therefore, it is less likely to find another peak with high response to *colorful* aspect while summarizing other aspects (i.e., searching for peaks in other dimensions in latent space). This situation makes the K-means result look more pleasing not necessarily because it summarized more aspects with high quality, but because the aspect of *colorful* shows up repeatedly in k-means summarization. To support our argument, we examine top-5 responding images for each aspect found by LDADeep+ to read the meaning of each aspect. We found that aspect#5 (i.e., the fifth dimension in latent space learned by LDADeep+, shown in Fig. 3(c)) is likely to be the aspect of *colorful*. Throughout all images in searching result of Holi, the average response to the fifth dimension in LDADeep+ latent space is 0.093. In images of k-means summarization, it is 0.107 (slightly above the average) on average; LDADeep+ 0.051 (much less than the average). This indicates that LDADeep+ is actually avoiding repeated summarization of *colorful* aspect.



(a) K-means clustering result (b) LDADeep+ result

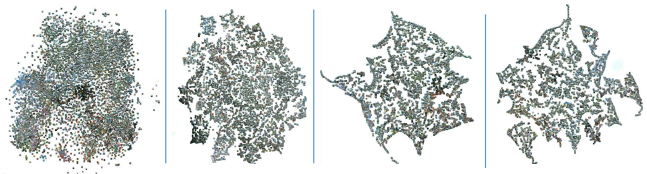
Fig. 2. Summarizations of Hanami. K-means method only focuses on finding the most diverse centroids (aspects) while LDADeep+ can emphasize both diversity and how frequently they are seen.



(a) K-means clustering result (b) LDADeep+ result (c) aspect#5 LDADeep+

Fig. 3. Summarizations of Holi. K-means method outperforms LDADeep+ on Holi case because LDADeep+ finds the aspect of *colorful* and avoids repeated summarization. On the contrary, k-means does not and people tend to favour colorful images.

3.2. Comparison of different representation of images



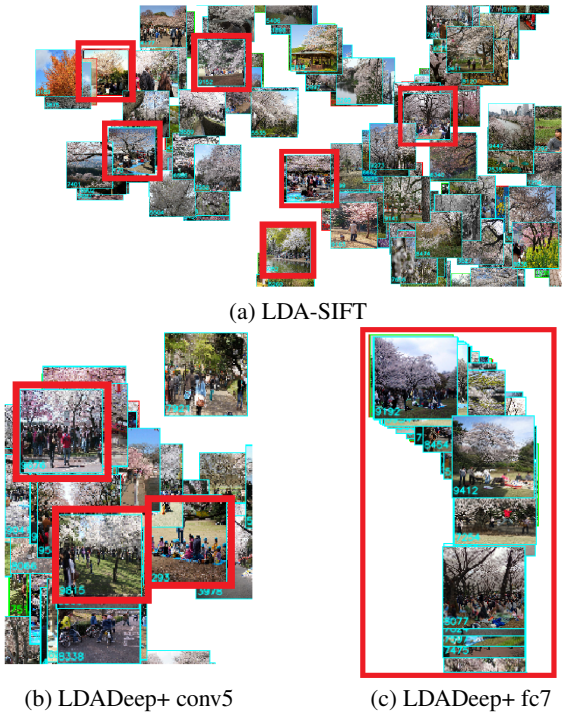
(a) (b) (c) (d)

Fig. 4. Visualization of clustering effect. Density of each result cluster implies the confidence of the model in recognizing this aspect; the clearer the boundaries between aspects are, the more discriminative the model is. (a) fc7 before aspect discovery; (b) LDA-SIFT; (c) LDADeep+ conv5; (d) LDADeep+ fc7

In the previous sections, we demonstrate that LDA model can achieve better summarization results. In this subsection, we discuss more about image representation with the comparison between LDADeep+ and LDA-SIFT. We visualize the relation between images by utilizing t-distributed stochastic neighbor embedding (t-SNE) [13] and conclude the rich semantic understanding ability LDADeep+ possesses. We further examine aspects discovered by LDADeep+ and LDA-SIFT to show that LDADeep+ learns sophisticated aspects by semantic level understanding of images while LDA-SIFT discovers aspects merely by simple pattern matching.

Fig. 4 shows several attributes of models: (1) density of each result cluster implies the confidence of the model in recognizing this aspect; (2) the clearer the boundaries between aspects are, the more discriminative the model is. In order to evaluate the quality of summarized aspects, we focus on a small portion of the summarization results in Fig. 5 to see how images with the same aspects are distributed in the aspect space. We pick a high-level concept, *picnic under cherry blossom tree*, which is a must-have activity when people go Hanami, as an example. In LDA-SIFT, images related to *picnic* do not get so close to each other and there are still images of other aspect, like *trees beside a lake* or *portrait of cherry blossom trees*, between them. The same situation appears in LDADeep+ model too if we use features of the fifth convolution (conv5) layer in AlexNet as the representation of images. As we apply our proposed LDADeep+ model with fc7 representation, the *picnic* aspect becomes easy to find and the region where images of *picnic* aspect lie in, namely, the bar-shaped region at the top of Fig. 4(d), is extraordinary clear: no images of other aspects come between them. This result highlights the discriminativeness of LDADeep+ model.

In order to understand those summarized aspects, we examine top-5 responding images for each aspect and find that LDADeep+ fc7 summarizes aspects in the highest conception level. Some examples of found aspects are shown in Fig. 6. Surprisingly, aspects that LDA-SIFT summarized tend to be low-level aspects of texture and pattern like *woven fabric*, *ruffle pattern* and *cherry blossom pattern*; LDADeep+ conv5 summarizes simple aspects but it misunderstands lawn as lake when summarizing *trees beside a lake*; for LDADeep+ fc7, it summarizes higher level aspects like *picnic*, *Japanese food*, *cherry blossom* and it summarizes *trees beside a lake* aspect without making mistakes. It is interesting to see that all models summarize *cherry blossom*. The *cherry blossom* LDA-SIFT summarizes is the fixed appearance; In contrast, LDADeep+ fc7 summarizes *cherry blossom* by its high-level concept. The angle and the distance are all different yet LDADeep+ model still recognizes them as the same aspect, moreover, cherry blossom is not even a category in ImageNet dataset which AlexNet is trained on. What surprises us most is that it learned the concept of *Japanese food*, which has a large variance in appearance. This finding points out how rich LDADeep+ is in terms of high-level knowledge and thus summarizes aspects with better quality and diverse results.



(a) LDA-SIFT

(b) LDADeep+ conv5 (c) LDADeep+ fc7

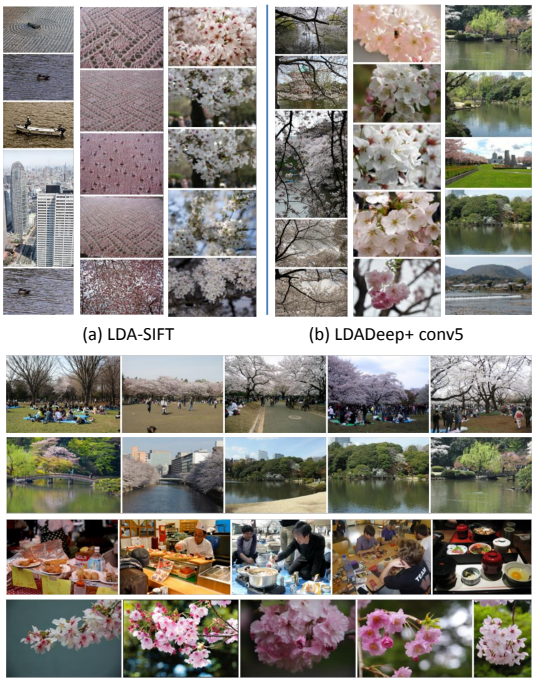
Fig. 5. Visualization of *picnic* aspect. We focus on a small portion (i.e., *picnic* aspect) of summarization results and mark images that are highly related to *picnic* with red boxes to see how the images of the aspect are distributed in aspect space. For LDA-SIFT there are still some images of other aspects that come between images of *picnic* aspect while for LDADeep+ fc7, it learns a clear aspect.

3.3. Application: Human attribute analysis

Last but not least, we apply LDADeep+ on one of the potential applications, the human attribute analysis task, as an example to show that by giving LDADeep+ some hints on what it should summarize, it can perform high quality aspect discovery and find out fine-grained aspects. Human action and activity are essential factors to many applications. We conduct our experiment on the searching result of keyword: Olympic 2012. By knowing what we should focus on, we first apply poselet [14], a body part detector, to find action patterns and only retain those results with high confidence to perform automatic aspect discovery. After this process, the aspects we attempt to discover get narrowed-down to focus on human attribute only. As shown in Fig. 7, the human actions found are meaningful and representative.

4. CONCLUSION

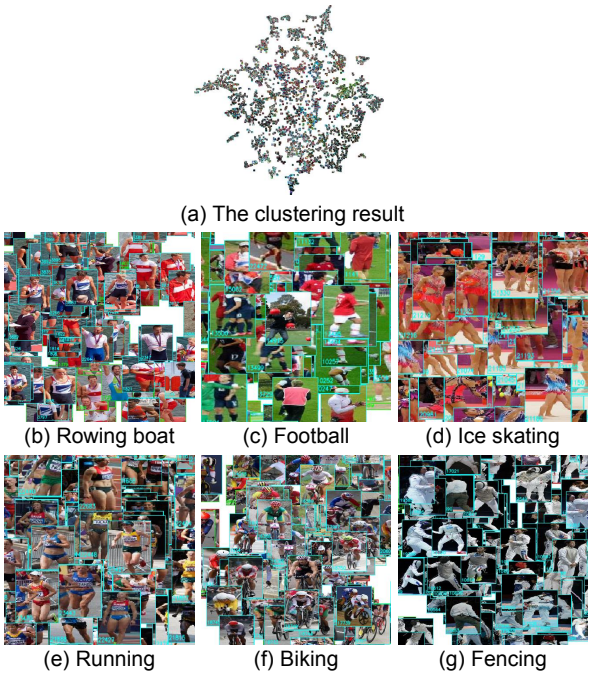
We propose a content-based automatic aspect discovery model: LDADeep+, which combines topic model with deep learning to learn high quality aspect, and solve the problem of automatic event summarization by discovering diverse and representative aspects with LDADeep+. The major break through is that we represent visual-word by employing deep learning global representation and build a LDA-based model on top of it to obtain high quality aspects. Topic model was once a very promising technique in computer vision application. However, with the great success deep learning made, it is worthwhile to think of whether there are other tasks that should be revisited and have a chance to make a breakthrough.



(a) LDA-SIFT (b) LDADeep+ conv5

(c) LDADeep+ fc7

Fig. 6. Example of found aspects. In (a) (b), aspects are shown in column-wise fashion and in (c), row-wise fashion. LDA-SIFT tends to discover low-level concepts (e.g., *woven fabric*, *ruffle pattern* and *cherry blossom pattern*) while LDADeep+ fc7 discovers high-level concepts (e.g., *picnic*, *Japanese food*, *cherry blossom*, and *trees beside a lake*) that focus on the semantic meaning of photos.



(a) The clustering result

(b) Rowing boat (c) Football (d) Ice skating

(e) Running (f) Biking (g) Fencing

Fig. 7. By preprocessing with poselet, LDADeep+ discovers aspects of Olympic Game related to human attributes like clothes and poses without specifying any target sub-event names.

5. REFERENCES

[1] David M. Blei and Andrew Y. Ng, "Latent dirichlet allocation," *JMLR.*, 2003.

[2] Rachit Arora and Balaraman Ravindran, "Latent dirichlet allocation based multi-document summarization," in *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*. 2003, ACM.

[3] Ana PB Lopes and Sandra EF de Avila, "A bag-of-features approach based on hue-sift descriptor for nude detection," in *Signal Processing Conference*, 2009.

[4] Xiaoli Yuan, Jing Yu, and Qin, "A sift-lbp image retrieval model based on bag of features," in *ICIP*, 2011.

[5] Bryan C. Russell, William T. Freeman, and Efros, "Using multiple segmentations to discover objects and their extent in image collections," in *CVPR*. 2006, IEEE Computer Society.

[6] Pulkit Agrawal, Ross Girshick, and Jitendra Malik, "Analyzing the performance of multilayer neural networks for object recognition," in *Computer Vision–ECCV 2014*. Springer, 2014.

[7] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, 2008.

[8] Christian Szegedy and Wei Liu, "Going deeper with convolutions," *CoRR*, 2014.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[10] Bart Thomee, David A. Shamma, and Friedland, "The new data and new challenges in multimedia research," *arXiv preprint arXiv:1503.01817*, 2015.

[11] Jia Deng, Wei Dong, and Socher, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.

[12] Ali S Razavian, Hossein Azizpour, and Sullivan, "Cnn features off-the-shelf: an astounding baseline for recognition," in *CVPRW*. IEEE, 2014.

[13] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *JMLR*, 2008.

[14] Lubomir Bourdev and Jitendra Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *Computer Vision*. IEEE, 2009.