Journal of Computer and Communications, 2014, 2, 54-60
Published Online July 2014 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2014.29008

Scientific Research

Learning Actions from the Identity in the Web

Khawla Hussein Ali, Tianjiang Wang

Huazhong University of Science and Technology (HUST), Wuhan, China
Email: khawlahussein@yahoo.com

Received April 2014

Abstract

This paper proposes an efficient and simple method for identity recognition in uncontrolled videos. The idea is to use images collected from the web to learn representations of actions related with identity, use this knowledge to automatically annotate identity in videos. Our approach is unsupervised where it can identify the identity of human in the video like YouTube directly through the knowledge of his actions. Its benefits are two-fold: 1) we can improve retrieval of identity images, and 2) we can collect a database of action poses related with identity, which can then be used in tagging videos. We present the simple experimental evidence that using action images related with identity collected from the web, annotating identity is possible.

Keywords

Action Recognition, HOG, SVM Classification

1. Introduction and Related Woks

Action recognition "in the wild" is often a very difficult problem for computer vision.

When the camera is non-stationary, and the background is fairly complicated, it is often difficult to infer the foreground features and the complex dynamics that are related to an action. Moreover, motion blur, serious occlusions and low resolution present additional challenges that cause the extracted features to be largely noisy [1]. Most research in human action recognition to date has focused on videos taken in controlled environments working with limited action vocabularies. Standard datasets, like KTH and Weizmann, formed for this purpose are well-explored in various studies, e.g. [2]-[6] and many more. However, real world videos rarely exhibit such consistent and relatively simple settings. Instead, there is a wide range environment where the actions can possibly take place, together with a large variety of possible actions that can be observed. Towards a more generic action recognition system, we propose to "learn" action representations from the web related with an identity and while doing this, improve the precision of the retrieved action images. Recent works [5] [7] [8] show that action recognition based on key poses from single video frames is possible. However, these methods require training with large amounts of video, especially if the system is to recognize actions in real world videos. Finding enough labeled video data that covers a diverse set of poses is quite challenging.

Internet contain rich sources of information, where one person may have several images from the Internet

How to cite this paper: Ali, K.H. and Wang, T.J. (2014) Learning Actions from the Identity in the Web. *Journal of Computer and Communications*, **2**, 54-60. http://dx.doi.org/10.4236/jcc.2014.29008

taken from different viewpoints, different clothes and may not be alone is present in the image, it may be with other people or other things, so our work is to focus on one person with single action and distinguish to learn his actions for only simple actions such as running, walking, clapping. So later we annotate the identity of the person in the video like YouTube, by learning his actions from the web.

Our work extends Nazli *et al.* [1] and continuous to join two lines of research "internet vision" and "action recognition" together.

Based on the assumption that the set of retrieved images contains relevant images of the queried identity, we construct a dataset of action images in an incremental manner. This yields a large image set especially if the person is famous, he has many images in the Web, which includes images of actions taken from multiple viewpoints in a range of environments, performed by that person who have different clothing. The images mostly present the "key poses" since these images try to convey the action with a single pose. There are challenges that come at the expense of this broad and representative data. First, the retrieved images are very noisy, since the web is very diverse. For example, for an "Abd-Rahman walking" query, a search engine is likely to retrieve images of along with that query, so maybe unnecessary and noisy. Our method must perform well in the presence of such diverse. We use the resulting dataset to annotate identity in videos of uncontrolled environments, like YouTube videos. Models trained with image data, of course, will be inferior to action models trained on videos solely; however, these models can serve as a basis for pruning the possible set of actions in a given video. In this work, we restrict our domain to actions which have characteristics that can be identifiable from a single monocular image, such as "running", "walking", "hand waving", "clapping". Our main contributions are:

- Addressing the problem of action image related with identity retrieval and proposing a system which incrementally collects them from the webby simple text querying;
- Building action models by using the noisy set of images in an unsupervised fashion; and
- Using the models to annotate human identity in uncontrolled videos.

Our method first collects an initial image set for the identity by querying the web. For the initial set of images retrieved for the identity, we fit a Bayesian Content-Based Image Retrieval instead of logistic regression classifier that used by Nazli *et al*. [1] to discriminate the foreground features of the related action of that identity from the background. Using this initial classifier, we then incrementally collect more action images related with identity and, at the same time, refine our model. This iterative process yields a more "cleaned" image set for that identity where a good many of the non-relevant images are removed. We then train separate local action classifiers and use these classifiers to annotate the identity in videos. Content-based image retrieval research has focused primarily on the retrieval of certain (and mostly unarticulated) objects; see [9]-[14], for some recent work. These works mostly rely on the knowledge of object classes and generic feature extraction. In our knowledge, only one work, Nazli *et al.* [1] have dealt with the retrieval of action images from the web, and no work with retrieval action recognition that related with identity. Moreover, while these works provide methods to collect datasets from the web, the final datasets are not mostly leveraged for further tasks. We accomplish this by making use of the collected dataset in a separate real-world domain. Little work has been done with generic videos like YouTube videos, where the resolution is low and the recording environment is nonuniform. Zanetti, *et al.* [12] recently noted the challenges of working with web videos. Niebles, *et al.* [15] present a method to detect moving people in such videos. Tran, *et al.* [6] detects actions in YouTube Badminton videos with fairly static backgrounds. Our method shown in **Figure 1** is applicable to videos with a broader range of settings. All previous methods are highly dependent on image details to extract features such as faces or body parts [16].

2. Image Representation and Building Action Models

To begin, we are given the initial results of a keyword query to an image search engine. For each of the retrieved images, we first extract the location of the human(s). If no humans are detected in an image, then that image is discarded. We use the implementation of Felzenswalb *et al*.'s human detector [17], which has been shown to be effective to detect human as in **Figure 2**. However, the background regions contained in the boxes do not provide any information about a specific person. In fact, when the same person is sighted indifferent surroundings, the background context causes ambiguity. Several segmentation methods could be used to separate the regions of the background from the regions of the human; however, we found in our experiments that it is better to estimate one distribution for each of the background and the foreground using a kernel density estimator [18]-[22]. Assuming that the human's figure will be centered in the bounding box, we use the center points as initial samples

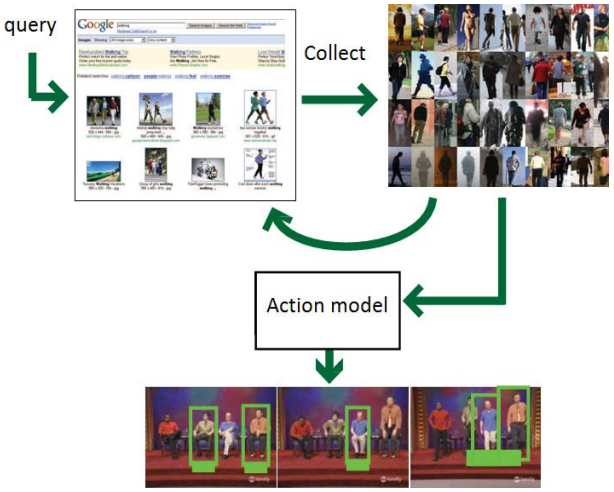
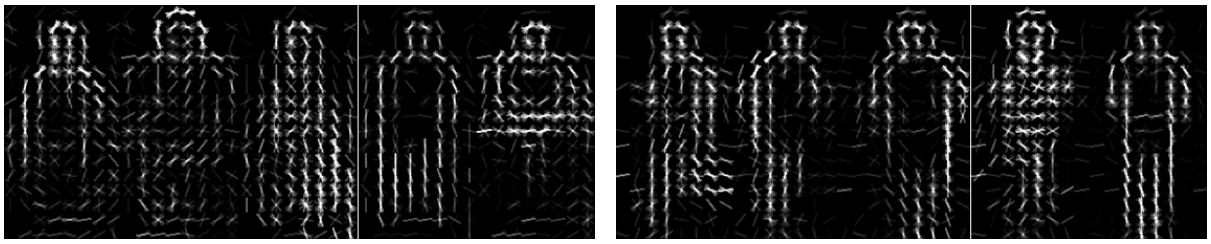


Figure 1. Illustrates our system. The system first gathers images by simply querying the name of the human on web image search engine like Google or Yahoo (Figure obtained from Nazli *et al.* [1]).



(a) (b)

Figure 2. Basis vectors of the actions found by HOG. (a) Running; (b) Walking.

to which eventually results in false matching. That it is better to estimate one distribution for each of the background and the foreground using a kernel density estimator [23] [24]. So we compute:

$$Score(x) = \frac{p(x|Dc)}{p(x)} \tag{1}$$

Assume a simple parameterized model, $P(x|\theta)$, and a prior on the parameters, $P(\theta)$. Since θ is unknown, to compute the score we need to average over all values of:

$$P(x) = \int p(x|\theta)\, p(\theta)\, \mathrm{d}\theta \tag{2}$$

$$P(x|Dc) = \int p(x|\theta)\, p(\theta|Dc)\, \mathrm{d}\theta$$

$$P(\theta|Dc) = \frac{p(Dc|\theta)\, p(\theta)}{p(Dc)} \tag{3}$$

Given a query of a small set of items, Bayesian Sets finds additional items that belong in this set. The score used for ranking items is based on the marginal likelihood of a probabilistic model. Bayesian Sets works well for finding more abstract set completions. Our training data is small, we use Baysian filter which have an advantage over (kNN or logistic regression), since the later will over fit. The Algorithm:

1) Input query word: *w* = "walking Abd-Rahman".

2) Find all training images with label *w*.

3) Take the binary feature vectors for these training images as query set and use Bayesian Sets algorithm.

For each image, x, in the unlabeled test set, we compute score(x) which measures the probability that x belongs in the set of images with the label w.

4) Return the images with the highest score. The algorithm is very fast [25].

After completion of the query, person detection, and feature extraction steps, we have a set of images that depict instances of the queried action plus a large number of irrelevant images, which includes images of other actions and noise images. The next task is to obtain a less noisy dataset that can serve as a training set in building our action model. We used the method of incremental learning in order to retrieve images of a person, and we assume the first ten images from the Web query images are images involving the person. But the images remain contain noise, we remove the noisy by adaptive filtering, by using a Wiener filter to the image adaptively, tailoring itself to the local image variance and can perform a smoothing.

3. Learning Classifiers for Action Discrimination

Using the above procedure incremental learning and using the techniques necessary, to get a clean database suitable for use in the classification. In this work, we have been used a framework for simultaneous tracking and action recognition using 2D part models with Dynamic Bayesian Action Network [17] [26]-[28]. The 2D part model allows more accurate pose alignment with the observations, thereby improving the recognition accuracy.

4. Recognizing Actions in Video

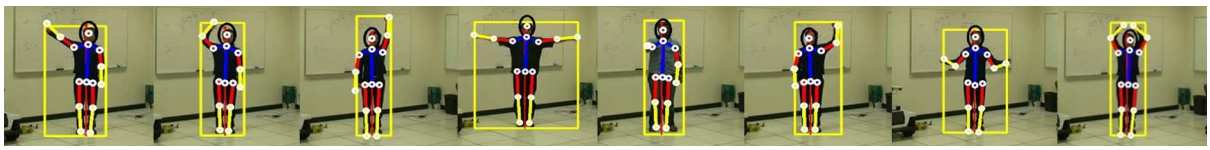
Having formed a dataset of action images and then learned classifiers, we want to annotate identity in videos. To do this, we first run the person detector [17] in each video frame. Once the humans have been detected, then recognition involves: perturbing the bounding box as in **Figure 3** and **Figure 4** to account for errors in localizing the humans, tracking of detections across frames.

5. Experimental Evaluation

We tested our approach on our dataset as in **Figure 5**; the video resolution is 960 × 540, and the height of person varies between 200 - 250 pixels across different persons.

5.1. Dataset Description

The identity dataset has 6 actions performed by 5 different persons, captured from a static camera in an outdoor



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3. Results on the gesture dataset: The bounding box shows the person position and the estimated pose is overlaid on top of the image, illustrated by limb axes and joints.

Figure 4. Example annotated frames from videos. We run the person detector [17] on these frames and create separate tracks for each person.

Figure 5. Example images from clusters of the identity "Abd-Rahman" and "Yang".

lab setting. The action set include-walking, running, jogging, hand-waving, hand-clapping, and boxing. Each action sequence in the dataset has exactly one person performing the action.

We evaluate our method in application of annotation of identity in YouTube videos. To collect the image dataset, we utilize query words related to identity on web search engines like Google and Yahoo Image Search. For querying each identity, in simplicity, we combine the action word (e.g. running) with identity like "Abd-Rahman". We collected images for six different actions: running, walking, sitting, hand waving, boxing and clapping. We use the dataset consists of YouTube videos related with the person that have considerably low resolution and moving cameras. This dataset has been used for person detection purposes and does not include identity annotations. We annotated 5 videos from this dataset, 150 frames in total, which includes the five actions in combination. Note that each video contains only one action, and since we will do frame by frame annotation.

5.2. Video Annotation

Our second experiment involves labeling the identity in videos by using the action models we form over web images. Besides our approach, for labeling identity in videos, we also tried two different classifiers: K-NN classifier and one-vs-all SVMs. In SVM classifier we use RBF kernels and are trained using bootstrapping in order to handle the noise [1] [25] [26]. We present the comparison of these techniques in **Table 1**. By the results, we observe that learning multiple local classifiers on poses is better than a single classifier for each action. **Figure 6** shows the confusion matrix for our method on this dataset. **Figure 7** shows the precision of collected images at recall level 20%. Most of the confusion occurs between running, jogging and walking actions. This is not surprising, since some of the walking poses involve a running pose for the legs, therefore some confusion is inevitable. This is the problem of composition of actions [29] and should be handled as a separate problem.

Table 1. Comparison of different classifiers and effects of smoothing on YouTube action annotations. The percentages shown are the average accuracies per frame.

Method	Accuracy
Ova SVM	90.3%
K-NN	89.2%
DBAN + 2D body parts	90.9%

	walk	run	jog	box	HWave	HClap
walk	**90.9**	0.0	9.1	0.0	0.0	0.0
run	0.0	**83.1**	16.9	0.0	0.0	0.0
jog	2.1	3.6	**94.3**	0.0	0.0	0.0
box	0.0	0.0	0.0	**85.2**	10.3	4.50
Hwave	0.0	0.0	0.0	0.0	**93.5**	6.5
Hclap	0.0	0.0	0.0	1.6	0.0	**98.4**

Figure 6. Per frame confusion matrix for identity annotation on YouTube videos related with actions.

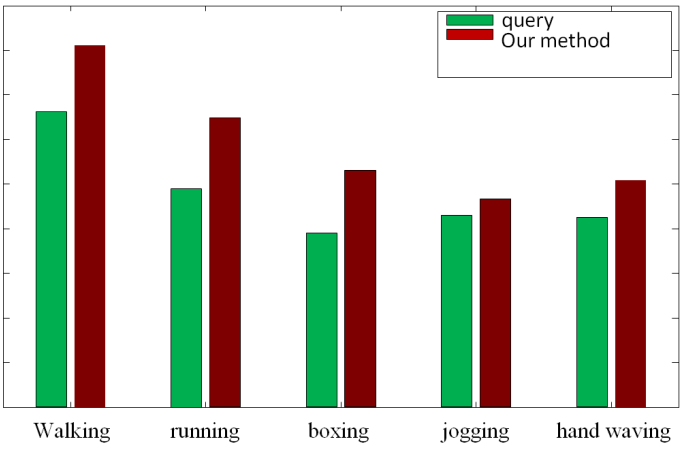


Figure 7. The precision of collected images at recall level 20%. Our method improves upon the precision of web image search in one identity of his actions.

6. Discussion and Conclusion

In this research, we have provided a simple and efficient fashion to distinguish the identity of the human in the video by learning his (here) actions from the Web. In this paper we assume a simple specification of the environment of the web image, carried out by the background as well as static and one action by one person. In the future we are trying to work on the identities from the web of real situations, like quite noisy, cluttered background and may be in the retrieval of the query more than one person. So the images must be more preprocessing before learning the action features. In the absence of a public datasets, we have tested the algorithm on our dataset. We can transfer the knowledge from the web images to annotate YouTube videos. Our system achieves promising results although many improvements can still be made. There is room for improvement, and extend this framework to recognize actions that include multiple persons and improving methods for dealing with noise and multi-modality and investigate for establishing a large dataset for this task.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments and suggestions that help to improve the quality of this manuscript.

References

[1] Ikizler-Cinbis, N. and Sclaroff, S. Object, Scene and Actions: Combining Multiple Features for Human Action Recognition.

[2] Oreifej, O., Mehran, R. and Shah, M. (2010) Human Identity Recognition in Aerial Images. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[3] Heller and Ghahramani, (2006) Bayesian Content-Based Image Retrieval.

[4] Singh, V.K. and Nevatia, R. Human Action Recognition Using a Dynamic Bayesian Action Network with 2D Part Models.

[5] Felzenszwalb, P., McAllester, D. and Ramanan, D. (2008) A Discriminatively Trained, Multiscale, Deformable Part Model. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[6] Natarajan, P., Singh, V.K. and Nevatia, R. (2010) Learning 3D Action Models from a Few 2D Videos for View Invariant Action Recognition. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[7] Kim, T.K. Wong, S.F., and Cipolla, R. (2007) Tensor Canonical Correlation Analysis for Action Classification. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[8] Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B. (2008) Learning Realistic Human Actions from Movies. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[9] Niebles, J.C., Han, B., Ferencz, A. and Li, F.-F. Extracting Moving People from Internet.

[10] Li, L.-J., Wang, G. and Li, F.-F. (2007) Optimol: Automatic Object Picture Collection via Incremental Model Learning. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[11] Martin, D., Fowlkes, C. and Malik, J. (2004) Learning to Detect Natural Image Boundaries Using Local Brightness, Color and Texture Cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**.

[12] Mikolajczyk, K. and Uemura, H. (2008) Action Recognition with Motion-Appearance Vocabulary Forest. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[13] Vijayanarasimhan, S. and Grauman, K. (2008) Keywords to Visual Categories: Multiple-Instance Learning for Weakly Supervised Object Categorization. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[14] Wang, G. and Forsyth, D. (2008) Object Image Retrieval by Exploiting Online Knowledge Resources. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[15] Niebles, J.C., Han, B., Ferencz, A. and Li, F.-F. (2008) Extracting Moving People from Web Videos. *European Conference on Computer Vision.*

[16] Niebles, J.C., Wang, H. and Li, F.-F. (2006) Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *British Machine Vision Conference.*

[17] Okada, R. and Soatto, S. (2008) Relevant Feature Selection for Human Pose Estimation and Localization in Cluttered Images. *European Conference on Computer Vision*.

[18] Wang, Y., Jiang, H., Drew, M.S., Li, Z.-N. and Mori, G. (2006) Unsupervised Discovery of Action Classes. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[19] Weinland, D. and Boyer, E. (2008) Action Recognition Using Exemplar Based Embedding. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[20] Zanetti, S., Zelnik-Manor, L. and Perona, P. A Walk Through the Web's Video.

[21] Gheissari, N., Sebastian, T. and Hartley, R. (2006) Person reidentification Using Spatiotemporal Appearance. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[22] Anguelov, D., Lee, K.-C., Gokturk, S. and Sumengen, B. (2007) Contextual Identity Recognition in Personal Photo Albums. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[23] Schindler, K. and van Gool, L. (2008) Action Snippets: How Many Frames Does Human Action Recognition Require? *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[24] Schroff, F., Criminisi, A. and Zisserman, A. (2007) Harvesting Image Databases from the Web. *International Conference on Computer Vision*.

[25] Schuldt, C., Laptev, I. and Caputo, B. (2004) Recognizing Human Actions: A Local SVM Approach. *International Conference on Pattern Recognition* (*ICPR*).

[26] Thurau, C. and Hlavac, V. (2008) Pose Primitive Based Human Action Recognition in Videos or Still Images. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

[27] Andreas Christmann1 and Robert Hable. On the Bootstrap Approach for Support Vector Machines and Related Kernel Based Methods.

[28] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annuals of Statistics*, **7**, 1-26.

[29] Tran, D. and Sorokin, A. (2008) Human Activity Recognition with Metric Learning. *European Conference on Computer Vision*.