Computer Vision Lab SS16
P-CNN: Pose-based CNN Features for Action Recognition

Ahmed Abdelbaki
University of Bonn
Bonn, Germany
abdelbak@iai.uni-bonn.de

Abstract

This work targets human action recognition in video based on the state of art deep convolutional neural network (CNN). Recent methods represents actions by statistics of local video features,here we argue for the importance of a representation derived from human pose. We are using an approach called Pose-based Convolutional Neural Network descriptor (P-CNN) for action recognition. The descriptor aggregates motion and appearance information through patches of human body parts. We investigate different CNN model fro the P-CNN approach and experiment P-CNN features obtained both for automatically estimated and manually annotated human poses. We evaluate our method on the recent and challenging JHMDB dataset. Our evaluation shows consistent improvement over the original approach.

1. Introduction

Recognition of human actions is an important step toward fully automatic understanding of dynamic scenes. Despite significant progress in recent years, action recognition remains a difficult challenge. Common problems stem from the strong variations of people and scenes in motion and appearance. Other factors include subtle differences of fine-grained actions, for example when manipulating small objects or assessing the quality of sports actions.

The majority of recent methods recognize actions based on statistical representations of local motion descriptors [11, 7, 13]. These approaches are very successful in recognizing coarse action (standing up, hand-shaking, dancing) in challenging scenes with camera motions, occlusions, multiple people, etc. Global approaches, however, are lacking structure and may not be optimal to recognize subtle variations, e.g. to distinguish correct and incorrect golf swings or to recognize fine-grained cooking actions

In this work, we are using the P-CNN approach which is a new action descriptor based on human poses. Provided with tracks of body joints over time,it combines motion and appearance features for body parts. In [2], they explore CNN features obtained separately for each body part in each frame due to the recent success of CNN [5, 8]. We use appearance and motion-based CNN features computed for each track of body parts and investigate different patch extraction mechanisms which are based on human joints pose. Moreover, we examine different CNN architecture for the feature extraction by comparing between the ”VGG-F” Net [1] and Microsoft Residual Networks(ResNet) [3]. The CNN features extraction pipeline of (P-CNN) is illustrated in figure 1.

The rest of the paper/technical report is organized as follows. Related work is discussed in Section2. Section3 introduces the P-CNN approach and CNN features extraction. Our contribution of the improved P-CNN is explained in Section4. We present the Dataset used in our experiments in Section5 and summarize the experimental results in Section6.

2. Related work

Action Recognition in the last decade has been dominated by local and hand crafted features. In particular, Dense Trajectories (DT) features[13] combined with Fisher Vector (FV)[10] aggregation have shown magnificent results for many challenging benchmarks.

The latest breakthroughs in Convolutional Neural Networks (CNN) have impacted a significant progress in image classification and other vision tasks.Although the applications of CNNs to action recognition in video[12, 14] has shown only limited improvements so far. P-CNN approach extend previous global CNN methods and address action recognition using CNN descriptors at the local level of human body parts.

Most of the recent methods for action recognition deploy global aggregation of local video descriptors. Such representations provide invariance to numerous variations

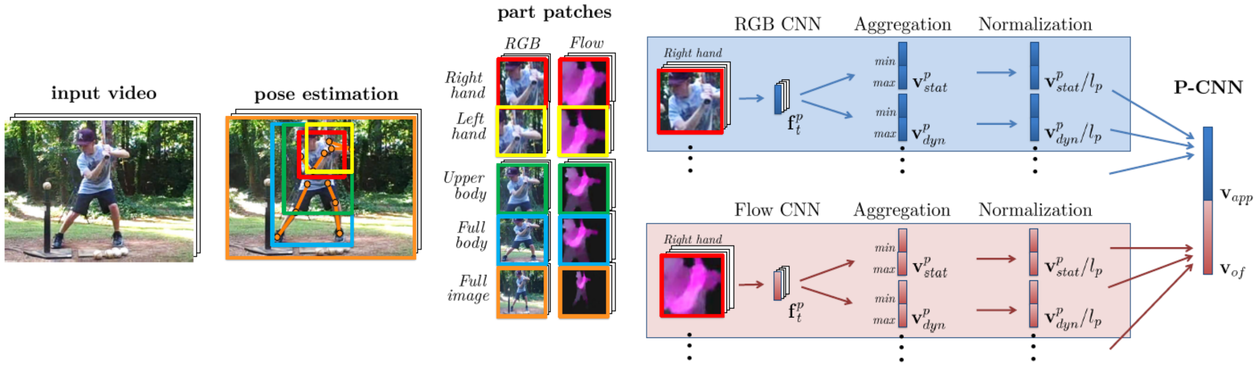

Figure 1. P-CNN features. From left to right: Input video. Human pose and corresponding human body parts for one frame of the video. Patches of appearance (RGB) and optical flow for human body parts. One RGB and one flow CNN descriptor f_t^p is extracted per frame t and per part p (an example is shown for the human body part *right hand*). Static frame descriptors f_t^p are aggregated over time using *min* and *max* to obtain the video descriptor v_{stat}^p. Similarly, temporal differences of f_t^p are aggregated to v_{dyn}^p. Video descriptors are normalized and concatenated over parts p and aggregation schemes into appearance features v_{app} and flow features v_{of}. The final P-CNN feature is the concatenation of v_{app} and v_{of}.

in the video but may fail to capture important spatio-temporal structure. For fine-grained action recognition, previous methods represented person-object interactions by joint tracking of hands and objects[9] or, by linking object proposals[15]. Alternative methods represent action using positions and temporal difference of body joints. Hence, it seems that Pose Estimation is a very important task for action recognition;however, reliable human pose estimation is still a challenging task. A recent study [4] reports significant gains provided by dynamic human pose features in case of the availability of reliable pose estimation. P-CNN extend the work [4] and designed a new CNN-based representation for human actions combing positions, appearance and motion of human joints.

3. P-CNN: Pose-based CNN features

Compare the following:

`$conf_a$` $conf_a$
`$\mathit{conf}_a$` conf_a

See The TeXbook, p165.

The space after *e.g.*, meaning "for example", should not be a sentence-ending space. So *e.g.* is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using "et alia", shortened to "*et al.*" (not "*et. al.*" as "*et*" is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct: " Frobnication has been trendy lately. It was introduced by Alpher [**?**], and subsequently developed by Alpher and Fotheringham-Smythe [**?**], and Alpher *et al.* [**?**]."

This is incorrect: "... subsequently developed by Alpher *et al.* [**?**] ..." because reference [**?**] has just two authors. If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al.*

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [**?**, **?**, **?**] to [**?**, **?**, **?**].

4. IPCNN: Improved P-CNN

All text must be in a two-column format. The total allowable width of the text area is $6\frac{7}{8}$ inches (17.5 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1.0 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

5. Datasets

In our experiments, we use a well-known dataset called JHMDB (Joint-annotated Human Motion Data Base)[4]. We present it in the following.

JHMDB is a subset of **HMDB**[6], see figure 2. It contains 21 human actions, such as *brush hair, climb, golf, run or sit*. Video clips are restricted to the duration of the action. There are between 36 and 55 clips per action for a total of 928 clips. Each clip contains between 15 and 40 frames of size 320 X 240. Human pose is annotated in each of the 31838 frames. There are 3 train/test splits for the JHMDB

Figure 2. Examples from JHMDB dataset

dataset and evaluation averages the results over these three splits. The metric used is accuracy: each clip is assigned an action label corresponding to the maximum value among the scores returned by the action classifiers.

6. Experimental results

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title 1-3/8 inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Figure and table captions should be 9-point Roman type as in Figures **??** and **??**. Short captions should be centred.

Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1. Results. Ours is better.

6.1. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

6.2. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [**?**]. Where appropriate, include the name(s) of editors of referenced books.

6.3. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

References

[1] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[2] G. Chéron, I. Laptev, and C. Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE*

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

International Conference on Computer Vision, pages 3218–3226, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[4] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3192–3199, 2013.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[9] B. Ni, V. R. Paramathayalan, and P. Moulin. Multiple granularity analysis for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 756–763, 2014.

[10] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010.

[11] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[12] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[13] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[14] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702, 2015.

[15] Y. Zhou, B. Ni, R. Hong, M. Wang, and Q. Tian. Interaction part mining: A mid-level approach for fine-grained action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3323–3331, 2015.