ACTION RECOGNITION USING SALIENT NEIGHBORING HISTOGRAMS

Ren, Huamin; Moeslund, Thomas B.

Published in:
IEEE International Conference on Image Processing

Publication date:
2013

Document Version
Accepted author manuscript, peer reviewed version

ACTION RECOGNITION USING SALIENT NEIGHBORING HISTOGRAMS

Huamin Ren, Thomas B. Moeslund

Visual Analysis of People Lab, Aalborg University, Denmark

ABSTRACT

Combining spatio-temporal interest points with Bag-of-Words models achieves state-of-the-art performance in action recognition. However, existing methods based on "bag-of-words" models either are too local to capture the variance in space/time or fail to solve the ambiguity problem in spatial and temporal dimensions. Instead, we propose a salient vocabulary construction algorithm to select visual words from a global point of view, and form compact descriptors to represent discriminative histograms in the neighborhoods. Those salient neighboring histograms are then trained to model different actions. Our approach yields a competitive result on the KTH dataset compare to state-of-the-art methods. On the more challenging UCF Sports dataset, we obtain 95.21%, which is approximately 4% better than the current best published results.

Index Terms— Salient visual words, neighboring histograms, action recognition

1. INTRODUCTION

Automatic action recognition is helpful in applications such as video surveillance, content-based video summarization, interactive computer applications etc. It gains more and more attention in recent years with the advance and prevalence of low-cost low-power sensors, computing devices and networks [1] [2]. The goal of human action recognition is to automatically analyze ongoing activities from an unknown video. Researchers from different application domains have investigated action recognition for the past decades and developed a diversity of approaches and techniques: some researchers seek ways to monitor the actor's movements in the suited environment directly by tracking, body pose estimation etc [3] [4], while others try to categorize actions based on the overall pattern in the video [5] [6] [7] [8]. Among them, spatio-temporal interest points, combined with bag-of-words models have achieved state-of-the-art results for action recognition [6] [9] [10].

Extracting spatio-temporal local features and then quantizing them into bag-of-words representations have several advantages: background substraction is not required, the descriptors are scale and rotation invariant in most cases, which are particularly suitable for recognizing periodic actions. Cuboid was first proposed in [5] as a spatio-temporal feature detector for human action recognition. They obtained a sparse distribution of interest points from a video, then associated a small 3-D volume (cuboid) to each interest point, which captured pixel appearance values of the interest point's neighborhood in the space and time. A library of cuboid prototypes was constructed by clustering cuboids appearance with K-means clustering. As a result, each action was modeled as a histogram of cuboid types detected in 3-D space-time volume. Due to the success of cuboids, various spatio-temporal feature extractors have been developed based on cuboids. Liu et al. presented a methodology to prune cuboid features to choose important and meaningful features [6]. Bregonzio et al. later proposed an improved detector for extracting cuboid features [11]. Despite of the success these approaches have achieved, there is also a key limitation which has been pointed out by several researchers: these spatio-temporal local interest point representations can be too local and fail to capture adequate spatial or temporal relationships [8].

Trajectory-based methodologies track features according to their spatial and temporal variations, which are capable of incorporating temporal information of the same feature in a certain period. In [12], the authors extracted feature trajectories by tracking Harris3D interest points [13] with the KLT tracker [14], and represented trajectories as sequences of log-polar quantized velocities. Sun et al. [15] extracted trajectories by matching SIFT descriptors between consecutive frames. Wang et al. [16] proposed an efficient method to extract dense trajectories by using dense sampling [17], which showed promising results over sparse interest points for action recognition in a recent evaluation [17]. Despite of the progress, the computation cost in tracking features is always expensive and therefore infeasible for large video datasets. Comparatively, matching features by modeling spatio-temporal relationships, or creating descriptors from neighboring spatio-temporal features can be good solutions. Ryoo and Aggarwal [18] introduced a novel spatio-temporal relationship matching to measure structural similarity between sets of features from two videos, which was thereby able to detect and localize complex non-periodic activities. Zhang et al. [19] proposed an approach called "spatio-temporal phrase", which encoded rich temporal ordering and spatial geometry information of local words. Kovashka

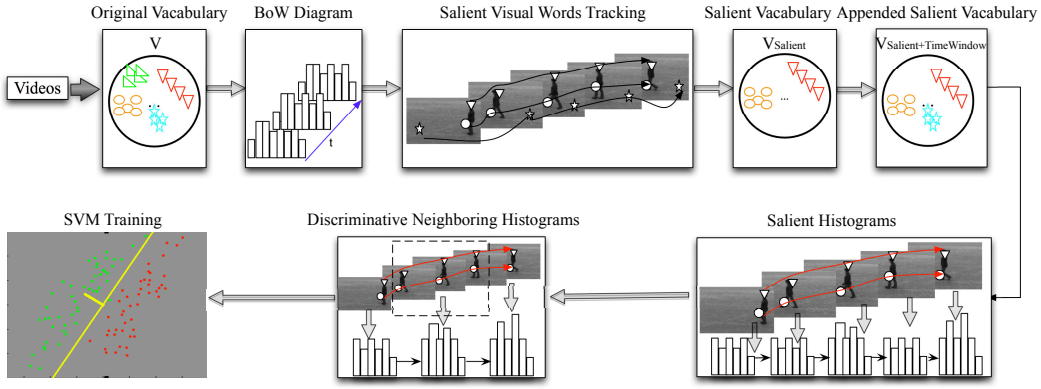

Fig. 1. Overview of the proposed system.

[8] proposed an algorithm to learn the shapes of space-time feature neighborhoods that were discriminative for a given action category. However, only creating visual words in a local space or time neighborhood cannot capture the long-term interactions of different body parts, which has a risk of incapability in recognizing the same action performed by different individuals. Moreover, the 2D space domain and 1D time domain in videos have very different characteristics, which put forward a new challenge: how to handle scaling ambiguity between the spatial and temporal dimensions when identifying nearest features to form trajectories or build space-time feature neighborhoods?

To address this problem, we propose a salient neighboring description algorithm, which builds a salient vocabulary from a global point of view in the temporal domain, and generates discriminative histograms from a neighboring point of view in the spatial domain. Our system for human action recognition can be explained explicitly in Figure 1. We apply salient visual words selection in Section 2 to choose representative visual words by tracing their variations across the video. These visual words turn out to have a high relationship with changes among consecutive frames in human actions. To address the "locality" problem in neighborhoods based algorithm, we enlarge the size of vocabulary by adding visual words existing in a predefined time window. Then we aggregate the neighboring frames into one group and represent each group as a discriminative neighboring histogram in Section 3. Finally, we regard these salient histograms as descriptors and train a multi-class support vector machine (SVM) to recognize actions.

2. GLOBALLY SALIENT VOCABULARY

An original vocabulary with the size of K is built after local features in videos are extracted. Then all the visual words are traversed to form histograms of the video V:

$$H(V) = H^0(V), H^1(V), ...H^m(V), ..., H^M(V), \quad (1)$$

where M is the number of frames in the video, $H^m(V)$ is the histogram descriptor of the m^{th} frame using Bag-of-Words models, which is a K-dimensional vector composed of $[H_1^m, H_2^m, ..., H_i^m, ...H_K^m]$.

Visual vocabulary, normally built by K-means clustering, always neglect the spatial or temporal relationship between local features in different frames. In action recognition where periodic behaviors are performed, temporal information gives an essential clue in action categories, e.g., the appearance of visual words representing the elbow joint in the human body is different in running compared to those in boxing. Accordingly, we adopt the salient visual words selection algorithm, which traces the variation of visual words in the entire video. Suppose we represent the trajectory of the i^{th} visual word as $H_i^0(V), H_i^1(V), ..., H_i^m(V)..., H_i^M(V)$, where m denotes the frame number. The variance of points (the i^{th} coordinate of the histogram representation) has a locally asymptotically stable property, mainly due to the facts that representative and consecutive visual information tend to draw close to specific extremes. Locally asymptotically stable points are defined as follows:

Definition 1: Suppose the i^{th} trajectory of the video V is represented as $\{H_i^0(V), H_i^1(V)...\}$. A point $H_i^{t^*}(V)$ is locally asymptotically stable at $t = t^*$ if it satisfies:

- $H_i^{t^*}(V)$ is stable and;
- $H_i^{t^*}(V)$ is locally attractive, e.g. there exists a δ such that:

$$\|t - t^*\| < \delta \Rightarrow \lim_{t \to t^*} H_i^t(V) = H_i^{t^*}(V) \quad (2)$$

The current visual word is regarded as "salient" if its trajectory contains locally asymptotically stable points. According to the second method of Lyapunov and Lyapunov's stability theory [20][21], the original system is locally asymptotically stable if we can find a Lyapunov-candidate-function. What we are doing here is somewhat the opposite, namely

given the assumption that the trajectory of a salient visual word is locally asymptotically stable and a Lyapunov-candidate-function exists, how to find out locally asymptotically stable points? A practical solution is to approximate the Lyapunov-candidate-function by using gradient function of the trajectory and detect local and stable points in it. Local points can be found by detecing extremes, stable points can be approximated by threshold filtering (δ_S), as defined in Eq. 3. If locally asymptotically stable points are detected in the gradient function, then the visual word is considered as "salient" and put into the salient vocabulary: $V_{Salient}$.

$$\delta_S = |H_i^{m+1}(V) + H_i^{m-1}(V) - 2 \cdot H_i^m(V)|. \quad (3)$$

However, the quantity of salient visual words is limited, thus histogram descriptors representing salient visual words frequency are very sparse, which leads to underfitting in classification. This problem can be addressed by tracking visual words from a time window and supplementing current salient vocabulary with those appearing in consecutive frames within the time window. However, determining the size of the time window is difficult, since automatic scene recognition is still a challenging field. For simplicity, we therefore use a fixed time window δ_T, and partition the video into scenarios. Visual words appearing continually in the current time window are appended into the vocabulary $V_{Salient}$. The new vocabulary is denoted as $V_{Salient+TimeWindow}$, and has the size: K_S ($K_S <= K$). Salient histograms of the video V can now be represented in Eq. 4, each element is a K_S-dimensional vector.

$$H_S(V) = H_S^0(V), H_S^1(V), ...H_S^m(V), ..., H_S^M(V). \quad (4)$$

3. DISCRIMINATIVE NEIGHBORING HISTOGRAMS

After selecting salient visual words, histogram descriptors of the video V - represented as $H_S(V)$ - are capable of indicating critical features in actions. But they are still not "discriminative" enough, which means histogram descriptors in consecutive frames may be too similar and affect the performance of the classification. To come up with discriminative descriptors, we generate neighboring histograms by partitioning frames with similar content into one group and generating a representative histogram description for the group.

By defining a parameter δ_N, which controls the changing range among consecutive frames, we can divide salient histograms of the video V into groups of neighboring histograms:

$$Group_m = [...H_S^{m-1}(V), H_S^m(V), H_S^{m+1}(V)...]. \quad (5)$$

The simplest way to calculate the descriptor for the neighborhood is by averaging, nevertheless, this approach doesn't perform well in our experiments. The variation within the group is normalized after averaging, which suppresses differentiation in the group. Correspondingly, we select the histogram with the largest variance as the representative descriptor. Thereafter, a video is represented as discriminative neighboring histograms, see Eq. 6 ($M' < M$).

$$H_N(V) = H_S^0(V), H_S^1(V), ...H_S^m(V), ..., H_S^{M'}(V). \quad (6)$$

In the final step, we train action models by using discriminative neighboring histograms and adopt a multi-class SVM with fast Intersection Kernels provided by [22] to classify different actions.

4. EXPERIMENTAL RESULTS

4.1. Datasets and Evaluation Details

We evaluate our approach on two benchmark datasets for human action recognition: the KTH dataset [23] and the UCF Sports dataset [24]. The KTH dataset contains a varied set of challenges including scale changes, variation in the speed of execution of an action, and indoor and outdoor illumination variations. The UCF Sports dataset includes a wide range of variations in viewpoint and scene background.

For the KTH dataset, we use the standard partition, dividing samples into training set (8+8 people) and test set (9 people). For the UCF Sports dataset, we test on each original sequence while training on all other sequences, following [23]. We train multi-class classifiers in two datasets and report the average accuaracy over all classes and compare our approach with the state-of-the-art.

4.2. Comparison to the state-of-the-art

KTH Dataset: We first compare our confusion matrix with the approach in [23], which introduced a template-based method to handle the locality problem in spatio-temporal interest points - based action recognition approaches. Diagonal elements in Fig. 2 with shadows indicate the action type in which our approach outperforms theirs. Our approach performs well in actions with obvious changes or fast movements, e.g., handclapping and boxing. Then we calculate the mean Average Precision (mAP) of our approach. We achieve a mAP of 94.78%, which is comparable to the approaches in [8] (94.53% as reported), the best accuracy we are aware of following the same setting as in [24].

UCF Sports Dataset: The advantage of our methods is even more obvious on UCF sports dataset, see Tab. 1. As far as we know, the accuracy of our method achieves the best result in per-class average recognition accuracy. Furthermore, we have observed that our discriminative neighboring descriptors perform well in fast moving action, e.g., skateboarding, as well as actions that last for a period. This is mainly due to the saliency of the vocabulary - visual words in a salient vocabulary represent visual information in the video either changing apparently, or lasting for a long period.

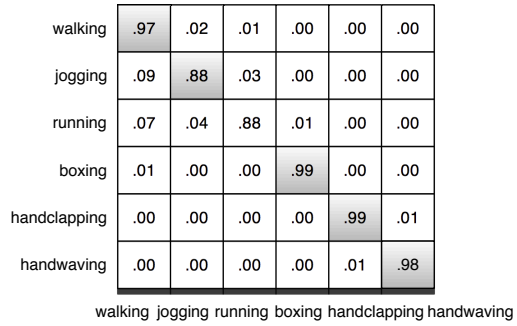

Fig. 2. Confusion matrix on KTH dataset.

Approach	Year	Accuracy/Class
Rodriguez et al. [23]	2008	69.2%
Varma et al. [25]	2009	85.2%
Wang et al. [17]	2009	85.6%
Kovashka et al. [8]	2010	87.27%
Kläser et al.[26]	2010	86.7%
Wang et al. [16]	2011	88.2%
Wu et al. [27]	2011	91.3%
Hara et al. [28]	2012	91.3%
Our method	**2013**	**95.21%**

Table 1. Comparative results on UCF Sports dataset.

4.3. Influencing Factors and Evaluations

First, we calculate the mAP over all action categories on KTH dataset, and compare the result on two vocabularies: $V_{Salient}$ and $V_{Salient+TimeWindow}$. The mAP for $V_{Salient}$ is 0.6043, while by tuning the value of δ_T, the mAP for $V_{Salient+TimeWindow}$ can be achieved as high as 0.9512. Here, the size of vocabulary K is set to 2000, and a randomly selection of 100,000 SIFT features are used to build the codebook on KTH. The success of complementing visual words lies in the fact that these visual information are helpful in classifying behaviors from different categories. For example, features representing thigh keep relatively stable in walking. When training models to classify features from walking to boxing, although these features are not salient in their respective videos, they are salient among different action categories. Thus adding these words as saliency helps to improve the performance.

Next, we change δ_N from a wide range to see the performance of neighboring histograms. Precision per class is calculated by adjusting δ_N. As can be seen from Fig. 3, the size of the time window can control the extent of "compactness" of neighboring histograms. The best result is achieved when δ_N is set to 150.

At last, we report the impact of vocabulary size on KTH dataset. We change the value of K from 500 to 5000, the best result is gained when K is set to 2000. We show some representative precision curves in Fig. 4. As can be seen, small vocabularies perform well in slow motion, while large ones are superior in fast movements.

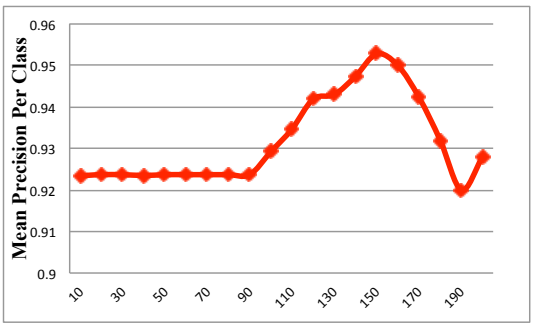

Fig. 3. Mean precision per class with different δ_N.

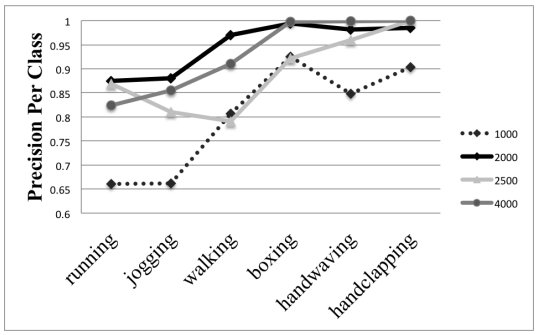

Fig. 4. Precision per class with different vocabularies.

5. CONCLUSIONS

We propose salient neighboring histogram representation to recognize actions in videos. Specifically, we introduced a globally salient vocabulary construction algorithm to pick out both representative visual words in the current video and discriminate information in the current action category. Additionally, we generated compact neighboring descriptions based on our salient vocabulary.

Our experiments on KTH and UCF Sports datasets demonstrate the success of introducing discriminative neighboring histograms into the already successful bag-of-words representation. In future work, we intend to combine spatio-temporal template matching with our salient neighboring histograms to classify activities from complicated backgrounds.

Acknowledgement
This work is supported by the Danish National Advanced Technology Foundation through the project: Managed Video as a Service (MVaaS).

6. REFERENCES

[1] J.K. Aggarwal and M.S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, pp. 16:1–16:43, apr 2011.

[2] P.K. Turaga, R. Chellappa, Subrahmanian V.S., and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 18, pp. 1473–1488, 2008.

[3] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," in *ICCV*. IEEE, 2009, pp. 925–931.

[4] L. Gorelick, M. Blank, E. Shechtman, M Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, pp. 2247–2253, 2007.

[5] P. Dollár, S. Rabaud, and V. Foster, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, 2005, pp. 65–72.

[6] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *CVPR*. IEEE, 2009, pp. 1996–2003.

[7] B. Chakraborty, M.B. Holte, T.B. Moeslund, and J. Gonzàlez, "Selective spatio-temporal interest points," *CVIU*, vol. 116, pp. 396–410, 2012.

[8] K. Kovashka, A.and Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *CVPR*. IEEE, 2010, pp. 2046–2053.

[9] M. Marszałek, J. Luo, I. Laptev, and C. Schmid, "Actions in context," in *CVPR*. IEEE, 2009, pp. 2929–2936.

[10] J. Niebles, H. Wang, and F. Li, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, vol. 79, pp. 299–318, 2008.

[11] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *CVPR*. IEEE, 2009, pp. 1948–1955.

[12] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *ICCV*. IEEE, 2009, pp. 104–111.

[13] I. Laptev and T. Lindeberg, "Space-time interest points," in *ICCV*. IEEE, 2003, pp. 432–439.

[14] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI*. William Kaufmann, 1981, pp. 674–679.

[15] J. Sun, X. Wu, S. Yan, L.F. Cheong, T. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *CVPR*. IEEE, 2009, pp. 2004–2011.

[16] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *CVPR*. IEEE, 2011, pp. 3169–3176.

[17] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*. BMVA Press, 2009, pp. 124.1–124.11.

[18] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *ICCV*. IEEE, 2009, pp. 1593–1600.

[19] Y. Zhang, X. Liu, M.C. Chang, W. Ge, and T. Chen, "Spatio-temporal phrases for activity recognition," in *ECCV*. Springer, 2012, pp. 707–721.

[20] A. M Liapunov, "Stability of motion," *Mathematics in science and engineering*, 1966.

[21] J.P. LaSalle and S. Lefschetz, *Stability by Lyapunov's Direct Method with Applications*, New York, Academic Press, 1961.

[22] S. Maji, C.B. Alexander, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *CVPR*. IEEE, 2008, pp. 1–8.

[23] M.D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR*. IEEE, 2008, pp. 1–8.

[24] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *ICPR*. IEEE, 2004, pp. 32–36.

[25] M. Varma and B.R. Babu, "More generality in efficient multiple kernel learning," in *ICML*. ACM, 2009, p. 134.

[26] A. Kläser, M. Marszałek, I. Laptev, and C. Schmid, "Will person detection help bag-of-features action recognition?," Tech. Rep., INRIA Grenoble - Rhône-Alpes, 2010.

[27] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *CVPR*. IEEE, 2011, pp. 489–496.

[28] S. Hara and B.A. Draper, "Scalable action recognition with a subspace forest," in *CVPR*. IEEE, 2012, pp. 1210–1217.