Human Action Classification Using N-Grams Visual Vocabulary

Ruber Hernández-García[1], Edel García-Reyes[2], Julián Ramos-Cózar[3], and Nicolás Guil[3]

[1] Digital Signals Departament, University of Informatics Sciences, Cuba
rhernandezg@uci.cu
[2] Pattern Recognition Department, Advanced Technologies Application Center, Cuba
egarcia@cenatav.co.cu
http://www.cenatav.co.cu/
[3] Dept. of Computer Architecture, University of Málaga, Spain
{julian,nico}@ac.uma.es
http://www.ac.uma.es/~vip

Abstract. Human action classification is an important task in computer vision. The *Bag-of-Words* model is a representation method very used in action classification techniques. In this work we propose an approach based on mid-level features representation for human action description. First, an optimal vocabulary is created without a preliminary number of visual words, which is a known problem of the *K-means* method. We introduce a graph-based video representation using the interest points relationships, in order to take into account the spatial and temporal layout. Finally, a second visual vocabulary based on n-grams is used for classification. This combines the representational power of graphs with the efficiency of the bag-of-words representation. The representation method was tested on the KTH dataset using STIP and MoSIFT descriptors and multi-class SVM with a chi-square kernel. The experimental results show that our approach using STIP descriptor outperforms the best results of state-of-art, meanwhile using MoSIFT descriptor are comparable to them.

Keywords: Human Action Classification, Bag-of-Words, Visual Words, Frequent Subgraphs, KTH dataset.

1 Introduction

Human action classification is an active research field in surveillance, human-computer interfaces, semantic video annotation, etc. *Bag-of-Words* (BoW) model is a representation method used in action classification approaches [12] [14]. The key idea is to quantize the high-dimensional space of local image descriptors to obtain a codebook of so-called visual words, sometimes also referred to as a visual vocabulary.

Visual words are sometimes justified on the basis of their ability to group semantically meaningful concepts, hence narrowing the semantic gap. However,

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 319–326, 2014.

one of the major challenges in BoW is the generation of visual vocabulary. The creation of visual vocabulary requires clustering many features descriptions detected in the training videos. *K-means* clustering and its variants are widely adopted. This method requires as parameter the vocabulary size, i.e. cluster number [19]. In addition, clustering is an unsupervised process and usually generates both descriptive and non descriptive words. Besides, another disadvantage of BoW model is that spatial and temporal constraints are ignored.

In this paper we present a mid-level features representation for human action description and its contribution is twofold. First, we propose to create a visual vocabulary using *K-means* and euclidean distance. The number of visual words is given in the range usually reported in the literature (i.e. 600 to 4000 words). The optimal vocabulary size is obtained applying *Leader* clustering algorithm with cosine similarity. Second, an intermediate representation of video is proposed, that combines the representational power of graphs with the efficiency of the bag-of-words representation. We extract the n-grade frequent subgraphs for each action class. Finally, a general visual vocabulary of n-grams is constructed from previous frequent subgraphs.

The experimental results show that our approach using STIP descriptor reports the best results of state-of-art (96.67%), meanwhile using MoSIFT descriptor are comparable to them (96.17%).

The rest of the paper is organized as follows. Section 2 briefly reviews visual vocabulary and graph-based representation for action categorization. Section 3 introduces the algorithm to obtain optimal vocabulary size and the n-grams visual codebook representation is presented. Experimental results using KTH Actions Database are given in Section 4. Finally, Section 5 concludes the paper.

2 Related Work

2.1 Visual Vocabulary Representation

The BoW model was introduced for action classification tasks by Laptev *et al.* [9] and Niebles *et al.* [12]. Local spatio-temporal features like Harris3D [8] or MoSIFT [3] are computed to obtain the video vocabulary, then a video is represented as a collection of visual words. A histogram representation is computed with the frequency of occurrence of every visual word in the vocabulary. Finally, any learning method can be used to classify them.

The visual words are obtained by clustering the interest points descriptors using *K-means* or any other unsupervised method. Unfortunately, these methods do not lead to an effective and compact vocabulary because many unnecessary and nondescriptive words are generated [19]. This can be alleviated by applying some ranking algorithm that sorts the words using some criterion of quality. In [4], the ranking algorithm proposed in [19] is applied to select Descriptive Visual Words (DVW) for action classification. This proves that a reduction in the vocabulary size can improve the classification accuracy. However, this reduction method requires a threshold for visual words selection and the dimension of vocabulary should be specified.

On the other hand, another disadvantage of BoW model is that spatial and temporal constraints are ignored. Therefore, losing spatio-temporal relationships is one of the important reasons that provoke the low descriptive power of classic visual words. Thus, some authors have proposed the use of correlograms to capture the co-occurrence of features [10]. These correlograms can be used to generate descriptive visual words [10] and visual phrases (bigrams) with the co-occurrence of two visual words [3] [19]. These works reported an improvement in the classification accuracy.

Visual bigrams capture the spatial and temporal relations between two visual words and presents better discriminative ability than the traditional codebook. However, it is possible to consider visual words in groups rather than pair, this could effectively capture the relationships among them. Moreover, n-grams represent high-level concepts and reduce the effects of image variations under different lighting conditions, views, scales and partial occlusion.

2.2 Graph-Based Representation

Several techniques have been developed to represent images in graph structure [1] [11] [13]. Graph-based representations provide powerful structural models, where the nodes can store local content and the edges can encode spatial information. Thus, Özdemir and Aksoy [13] propose an intermediate representation that combines the representational power of graphs with the efficiency of the bag-of-words representation.

The proposed approach in [13] represents each image with a histogram of frequent subgraphs where the subgraphs encode the local patches and their spatial arrangements. The subgraphs that are used as the visual words of the final bag-of-words model are selected using a frequent subgraph mining algorithm. The frequent subgraphs are used to avoid the need of fixing a complexity (in terms of the number of nodes). Consequently, the spatial structure in the image is encoded in a histogram, and the graph matching problem is transformed into a vector comparison that reduces the computational cost.

In video domain, the spatio-temporal relations between interest points can be exploited in order to built a graph as a video representation. The interest points detected in features description phase form the graph nodes, and their labels are determined from visual words assigned by BoW model. The graph is built by connecting every neighbor node pair with an edge. Two nodes are considered neighbors if the Euclidean distance between them is lesser than a threshold or selecting the nearest in neighborhood.

3 Proposed Method

3.1 Visual Vocabulary Reduction

The flowchart of our approach, depicted in Fig. 1, implements a mid-level representation with four levels of features representation. The first level of our representation model are the spatio-temporal interest points extracted from video.

In the second level we construct an initial visual vocabulary using classical *K-means* clustering algorithm. The third level of representation is obtained by re-clustering the codebook created in second level using *Leader* clustering algorithm [5] [17] with cosine similarity. We adopted this approach because this method is able to join the similar visual words in the same cluster reducing the vocabulary size to an optimal dimension.

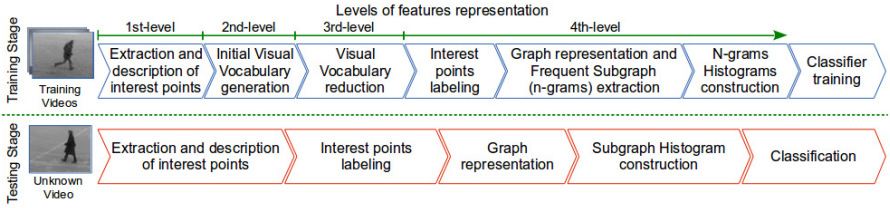

Fig. 1. Flowchart of our approach based on mid-level representation. In terms of knowledge discovering about visual aspects of a video: 1st-level represents the *"low-level visual features"*; 2nd-level correspond to *"visual words"* representation; 3rd-level groups similar visual words to extract more abstract *"visual concepts"*; and 4th-level construct *"visual n-grams"* in visual concepts co-relations terms.

Given a basic visual vocabulary $W = \{w_i\}_{i=1}^{k}$, where w_i is a visual word and k is the vocabulary size. Let C be a set of current clusters, Lv be the average value of calculated affinity matrix, Lvc_j be the average distance value in cluster j. We define the algorithm for optimal vocabulary construction that is presented in Algorithm 1.

Algorithm 1. Visual Vocabulary Reduction

Initialize $c_1 \in C$ and $w_1 \in c_1$. **Set** $iter \leftarrow 0$; $best\_iter \leftarrow 0$; $min\_criteria \leftarrow \infty$.
While $iter < maxiter$ **do**
  Randomize the order of W.
  For $\dot{w} \in$ *the rest* W **do**
    $S \leftarrow \min_{\dot{c} \in C}(similarity(\dot{w}, \dot{c}))$
    $j \leftarrow \arg\min_{\dot{c} \in C}(similarity(\dot{w}, \dot{c}))$
    If $((count(c_j) = 1$ *and* $S < Lv)$ *or* $(count(c_j) > 1$ *and* $S < Lvc_j))$
      $\dot{w} \in c_j$
      Update c_j centroid.
    Else
      $new\ c \in C$
      $\dot{w} \in$ *new* c
    End If
  End For
  Update C: $\forall\ \dot{w} \in \dot{c} \Rightarrow similarity(\dot{w}, \dot{c}) > 2 * Lvc_j$ re-locate $\dot{w}$.
  Calculate $Je_{iter} = Related\ Minimun\ Variance\ Criterion$ [5].
  $best\_iter \leftarrow \arg\min\,(Je_{iter}, Je_{best\_iter})$
End While

The first step in this algorithm is to put the first instance of visual words into a new cluster c_1. For a new visual word $\dot{w}$, we select the cluster c_i such as the similarity between the cluster centroid and $\dot{w}$ to be minimum. Then, we decide whether $\dot{w}$ is put inner c_i or in a new cluster based on that a data needs to be put inner new cluster if the distance to the current clusters is statistically big enough. Thus, we use Lvc and Lv as dynamical threshold. We adapt the algorithm in [17] updating the centroid when a new data is introduced in the cluster. Besides, to alleviate stability problem of leader clustering algorithm, we execute several clustering iterations with random order of data. At the end of each clustering iteration the partitions are updated and is selected the optimal partitioning with minimal variance criterion. Finally, the visual vocabulary with optimal size is returned.

3.2 N-Grams Visual Representation

The fourth level of our representation approach, depicted in Fig. 1, is obtained labeling the interest points from the reducing vocabulary created in third level of representation. Then we construct a k-Nearest Neighbour graph (k-NNG) for the interest points labeled. The n-grade frequent subgraphs are extracted for each action class. Finally, a general visual vocabulary of n-grams is constructed from previous subgraphs. Each video is represented with a histogram of frequent n-grams.

In order to limit the complexity of our graph representation a k-NNG is constructed with proximity distance threshold. We define a spatio-temporal distance d with the purpose to identify a interest point pair proximity [4]. Because interest points may have various scales, the value of d is computed with $d = scale_i \times D$, to achieve scale invariance, where $scale_i$ is the scale of the interest point and D is a parameter controlling the constraint of proximity relationships.

The graph mining literature includes several approaches for frequent subgraph mining. According to [18], frequent subgraph mining is to find every graph, g, whose support in a graph set, $GS = \{G_i | i = 1, \ldots, N\}$, is equal or greater than a threshold, $minSup$. The support of g in GS is denoted as $\sigma(g, GS)$, and is generally defined as the number of graphs in GS that have a subgraph which is isomorphic to g.

In our approach as in [13], the number of times that a subgraph occurs in a video provides more information than the knowledge about whether the subgraph is contained or not. Due to the graph set in our application includes many overlapping embeddings the Harmful overlap (HO) support [6] is utilized. Also, we use different and dynamic support thresholds for each class instead of a global threshold. The support threshold is equal to average support in the class. The sets of frequent subgraphs are found using the graph sets of training videos for each class independently. Finally, these sets of frequent subgraphs are joined into a single set S.

The n-grams histogram provides a powerful representation that is not as complex as full graph models, and reduces the complexity of graph similarity

computation [13]. The n-grams histogram for a video is constructed using the frequency of each subgraph from S in the video. Consequently, video can be classified in this feature space using multi-class support vector machine (SVM) with a chi-square kernel, as is described in testing stage in Fig. 1.

4 Experimental Results

In this section, we show the experimental results that validate the efficacy of the proposed approach for human action classification. The experiments were performed using the KTH actions dataset [14], because the interest points extracted on this database are related with action. In others actions datasets noising interest points are extracted on background, then previous filter process is necessary. Our main goal in this work is to prove the representative power of mid-level representation proposed. We follow the leave-on-out cross validation (LOOCV) evaluation criterion in our experimentation.

Each video sequence is represented as a bag of spatial-temporal features using STIP and MoSIFT descriptors. Descriptors are extracted running the implementations freely available on Internet for STIP[1] and MoSIFT[2] with the default parameters. First, initial visual vocabularies are constructed with different sizes (600, 1000, and 4000 visual words) using *K-means* clustering algorithm. We select subgraphs with degrees between 2 and 5. Additionally, a Support Vector Machine (SVM) has been utilized to classify actions to prove the entire performance of our approach.

Table 1 summarizes the most relevant results obtained. The first column shows the accuracy values obtained with initial vocabulary, whilst second and third columns present the accuracies using a vocabulary with optimal size and n-grams histograms respectively. As a general rule, accuracy is increased with our mid-level representation proposed. However, this tendency is more noticeable with STIP descriptor. It could be argued that the MoSIFT descriptor originates less redundant vocabularies. Best accuracies are reached for n-grams histograms representation. That proves the validity of our approach.

Table 1. Accuracies for the STIP and MoSIFT descriptors obtained with different representation levels in the KTH Actions Database. Best results are marked in bold.

STIP descriptor						MoSIFT descriptor					
Initial size	Acc	Optimal size	Acc	Number of n-grams	Acc	Initial size	Acc	Optimal size	Acc	Number of n-grams	Acc
600	94.64	201	94.82	1052	95.17	600	92.17	270	93.33	1286	94.86
1000	94.50	340	**95.32**	1549	**96.67**	1000	93.67	456	93.86	1833	**96.17**
4000	**95.99**	1188	95.15	2776	96.50	4000	**94.67**	1684	**94.83**	2989	95.90

[1] http://www.di.ens.fr/~laptev/download.html

[2] http://lastlaugh.inf.cs.cmu.edu/libscom/downloads.htm

Finally, Table 2 shows a comparison between our method and the performance reported by previous works using a similar experimental setup. Our approach using STIP descriptor reports the best results, meanwhile using MoSIFT descriptor our results are comparable to state-of-art.

Table 2. Comparison with other methods all using the KTH dataset with a LOOCV experimental set-up

Method	Avg. Accuracy
Our method (STIP, 1549 Frequent Subgraphs)	**96.67**
Chakraborty *et al.* (2012) [2]	96.35
Gao *et al.* (2010) [7]	96.33
Our method (MoSIFT, 1833 Frequent Subgraphs)	**96.17**
Chen and Hauptmann (2009) [3]	95.83
Wang *et al.* (2013) [16]	95.30
Thi *et al.* (2012) [15]	94.33

5 Conclusions

In this paper, we propose a new approach based on mid-level feature representations for human action description. We also present graph-based video representation using the interest points relationships. Our method faces some limitations of the BoW model. We create a vocabulary with optimal size without requires preliminary number of visual words. A representation using visual n-grams is proposed.

The experimental validation of mid-level feature representation presented here reached the best results in KTH database in comparisons with other methods using similar experimental setup. However, an evaluation with more complex human action video databases is still necessary in order to obtain concluding results.

As future work, we will label the graph edges to encode topological relationships between interest points to conserve a lot more semantics information. Also, its necessary to study the influence of subgraphs size on accuracy improvement. Besides, we think to increase levels of abstraction in the representation and the same time use sparse coding instead of *K-means* algorithm.

References

1. Acosta-Mendoza, N., Gago-Alonso, A., Medina-Pagola, J.E.: Frequent approximate subgraphs as features for graph-based image classification. Knowledge-Based Systems 27, 381–392 (2012)
2. Chakraborty, B., Holte, M.B., Moeslund, T.B., Gonzàlez, J.: Selective spatio-temporal interest points. Computer Vision and Image Understanding 116(3), 396–410 (2012)
3. Chen, M.Y., Hauptmann, A.: Mosift: Recognizing human actions in surveillance videos. Research Showcase 929, Carnegie Mellon University. School of Computer Science. Computer Science Department (2009)

4. Cózar, J.R., Hernández, R., Heredia, Y., González-Linares, J.M., Guil, N.: Reducing Vocabulary Size in Human Action Classification. In: Frontiers in Artificial Intelligence and Applications, vol. 243, pp. 1712–1719. IOS Press (2012)
5. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley Interscience (2001)
6. Fiedler, M., Borgelt, C.: Support computation for mining frequent subgraphs in a single graph. In: Proceedings of MLG-2007: 5th International Workshop on Mining and Learning with Graphs, pp. 1–6 (2007)
7. Gao, Z., Chen, M.-Y., Hauptmann, A.G., Cai, A.: Comparing Evaluation Protocols on the KTH Dataset. In: Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A. (eds.) HBU 2010. LNCS, vol. 6219, pp. 88–100. Springer, Heidelberg (2010)
8. Laptev, I., Lindeberg, T.: Space-time interest points. In: Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003), vol. 1, pp. 432–439 (2003)
9. Laptev, I., Marszaek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR0 2008, pp. 1–8 (2008)
10. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1–8 (2009)
11. Morales-González, A., García-Reyes, E.: Assessing the role of spatial relations for the object recognition task. In: Bloch, I., Cesar Jr., R.M. (eds.) CIARP 2010. LNCS, vol. 6419, pp. 549–556. Springer, Heidelberg (2010)
12. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning and of human and action categories and using and spatial-temporal words. International Journal on Computer Vision (79), 299–318 (2008)
13. Özdemir, B., Aksoy, S.: Image classification using subgraph histogram representation. In: ICPR 2010, pp. 1112–1115 (August 2010)
14. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 3, pp. 32–36 (August 2004)
15. Thi, T.H., Cheng, L., Zhang, J., Wang, L., Satoh, S.: Structured learning of local features for human action classification and localization. Image and Vision Computing 30(1), 1–14 (2012)
16. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision 103(1), 60–79 (2013)
17. Wiliem, A., Madasu, V.K., Boles, W.W., Yarlagadda, P.K.: Detecting uncommon and trajectories. In: Digital Image and Computing: Techniques and Applications, DICTA (December 2008)
18. Yan, X., Han, J.: gspan: Graph-based substructure pattern mining. In: IEEE International Conference on Data Mining, ICDM 2002, pp. 721–724 (2002)
19. Zhang, S., Tian, Q., Hua, G., Huang, Q., Li, S.: Descriptive visual words and visual phrases for image applications. In: Proceedings of the 17th ACM International Conference on Multimedia, pp. 75–84 (October 2009)