Visual Words Selection for Human Action Classification

Technical Report, May 2012

J.R. Cózar, J.M.[a] *González-Linares, N. Guil*

Dept. of Computer Architecture

University of Málaga

Málaga, Spain

{*julian, gonzalez, nico*}*@ac.uma.es*

R. Hernández, Y. Heredia

Departamento Señales Digitales

Universidad de las Ciencias Informticas

La Habana, Cuba

{*rhernandezg,yhernandezh*}*@uci.cu*

Abstract

Human action classification is an important task in computer vision. The Bag-of-Words model uses spatio-temporal features assigned to visual words of a vocabulary and some classification algorithm to attain this goal. In this work we have studied the effect of reducing the vocabulary size using a video word ranking method. We have applied this method to the KTH dataset to obtain a vocabulary with more descriptive words where the representation is more compact and efficient. Two feature descriptors, STIP and MoSIFT, and two classifiers, KNN and SVM, have been used to check the validity of our approach. Results for different vocabulary sizes show an improvement of the recognition rate whilst reducing the number of words as non-descriptive words are removed. Additionally, state-of-the-art performances are reached with this new compact vocabulary representation.

Keywords: Feature Selection and Extraction, Classification, Support Vector Machines, Computer Vision.

1 Introduction

Many applications in computer vision, such as surveillance, human-computer interfaces and semantic video annotation, are based on human action categorization. Thus, action recognition is an active research field in computer vision.

In the literature, two main approaches for human action recognition are used: holistic and part-based representations. The holistic representation focus on the whole body of the person, trying to search for characteristics such as contours or pose. On the other

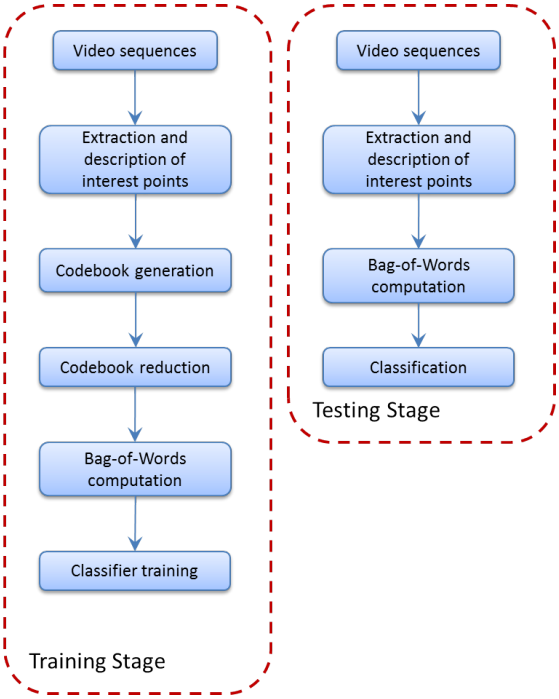

Fig. 1: Bag-of-Words framework. Training stage is applied to the training set to obtain a codebook and a classifier. Testing stage is applied to the test set using the previously computed codebook and classifier to obtain a label for every video.

hand, part-based representation consists of two steps: feature detection phase, in which space-time interest points are searched for in the video, and the feature description phase, in which a robust description of the area around them is computed and a model based on independent features (Bag-of-Words) or a model that can also contain structural information is built.

The creation of a vocabulary for a Bag-of-Words model requires clustering the features descriptions detected in the training videos. Clustering is an unsupervised process and usually generates both descriptive and non-descriptive words for the vocabulary. The aim of this paper is to study the influence of the selection of more descriptive words for vocabulary creation as regards action detection accuracy. As will be shown in the experimental results, accuracy is improved using reduced vocabularies.

To check the validity of our results we have tried different feature descriptors and classifiers. We have selected two well-known feature descriptors, STIP [1] and MOSIFT [2], and two popular classifiers, k-nearest neighbor (KNN) and support vector machines (SVM), in order to compare our results with other works.

The rest of the paper is organized as follows. Section 2 reviews visual vocabulary creation for action categorization. The algorithm for vocabulary selection is introduced in Section 3. Section 4 evaluates the impact of vocabulary reduction on the accuracy of action categorization. Finally, Section 5 concludes the paper.

2 Vocabulary Creation

The Bag-of-Words (BoW) model, depicted in Fig. 1, has been used in many multimedia related tasks such as image retrieval, object recognition or video event detection. With this model an image is represented as a collection of visual words that belong to a visual

vocabulary. This vocabulary is created in a training stage by clustering a large number of local feature descriptors. Then, a BoW representation is computed as a histogram with the frequency of occurrence of every visual word in the vocabulary, and finally, a Naïve Bayes classifier, a support vector machine (SVM), or any other learning method that has been previously trained in the training stage can be used for classification. The codebook of visual words and the classifier are computed only in the training stage.

One of the main disadvantages of this model is that spatial and temporal constraints are ignored, thus some authors as [3] have proposed the use of correlograms to capture the co-occurrence of features. This correlograms can be used to generate descriptive visual words and visual phrases that can be more effective in terms of efficiency and accuracy.

The BoW model has been also used for recognizing human actions. Local spatio-temporal features such as STIP or MoSIFT are computed to obtain the video vocabulary, and bigrams with the co-occurrence of two video words can be selected to enrich the vocabulary.

Space-Time Interest Points (STIP) is a feature detector proposed by Laptev in [1]. It extends the idea of the Harris interest point detector to the spatio-temporal domain. It builds a scale-space representation of the sequence by convolution with a spatio-temporal separable Gaussian kernel that has independent spatial and temporal variances. Then, a spatio-temporal second order matrix is computed and an extended version of the Harris corner function is used to detect the interest points. Different spatial and temporal scales can be used, and a Laplace operator can be applied over scales to compute scale-adapted space-time interest points. This method can capture the temporal extent of the features, allowing to distinguish between fast and slow movements.

These points are represented using histograms of oriented gradient and optical flow. For each point a volume proportional to the detection scales and centered at the detection point is subdivided in several cuboids, and for each cuboid a histogram of oriented gradient (HoG) and optical flow (HoF) is computed. These histograms are normalized and concatenated to form the interest point descriptor.

The MoSIFT algorithm [2] is another well-known method for extracting and describing interest points. It detects spatially distinctive interest points with substantial motion. It uses the SIFT algorithm to extract interesting points in the spatial domain, and selects those points with a 'sufficient' amount of optical flow around them. Later, these points are described using a combination of a histogram of optical flow (HoF) and a histogram of gradients (HoG).

The scale invariant SIFT points are computed using a scale-space kernel: a pyramid of Difference of Gaussians (DoG) images is obtained, and local extrema across adjacent scales are used as the interest points. Then, a multiple-scale optical flow pyramid is built over the DoG pyramid and those interest points with sufficient motion are selected. This method is not invariant to temporal scale, and consequently no motion constraint is imposed in the interest points.

After computing the descriptors of interest points, a codebook of visual words is generated. Visual words are obtained by clustering the interest points 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. This can be alleviated by applying a codebook reduction step where some ranking algorithm that sort the words by their descriptive quality can be used to select the most descriptive words. For example, in image retrieval, a ranking algorithm called *VisualWordRank* was proposed in [4] to select those words that appear more frequently in some visual category, and also co-occur with other frequent words in that visual category. These

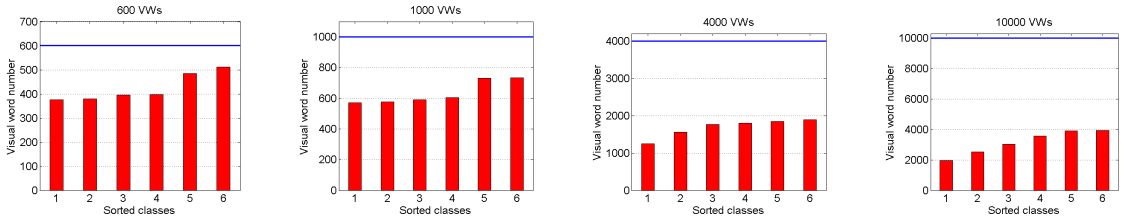

Fig. 2: Number of visual words used in each KTH action category for different STIP descriptor vocabulary sizes (600, 1000, 4000, 10000 VWs).

words, called Descriptive Visual Words (DVW), form a significantly lesser vocabulary and computational complexity is reduced as well. However, the vocabulary reduction presented in [4] is applied only to large databases of images. In this paper we will apply their method to human action categorization in videos and study the impact of vocabulary reduction on classification accuracy.

The codebook obtained in the previous step can be used to compute a histogram representation with the frequency of occurrence of every visual word in this vocabulary. These histograms are computed for every video to train a classifier such as a support vector machine (SVM). For multi-class classification the one-versus-all rule can be used to obtain a classifier for each separate class.

Finally, in the testing stage, the same feature extractor and descriptor is used in conjunction with the codebook computed in the training step to obtain a BoW representation that is classified using the trained SVM. The label of the classifier with the highest response is assigned to the test video and compared with the groundtruth to obtain accuracy values.

3 Vocabulary Reduction

It has been proven that the *K-means*-based clustering of visual words does not typically generate an effective and compact visual vocabulary. In addition, the descriptive power of visual vocabularies is influenced by the number of visual words used. The more visual words are extracted the better performance is achieved. However, the performance will be saturated when the number of words reaches certain levels [5]. Another consequence of using more visual words is that the relative number of visual words used in each category decreases. This behaviour is shown in Fig. 2, where the KTH action dataset and the STIP descriptor has been employed. This Figure depicts the number of different visual words used in the representation of different actions, sorted in ascending order. It is clearly shown that each category is described only with a portion of the total vocabulary. In other words, eliminating VWs from the vocabulary by selecting the more informative ones allows us to obtain a more efficient vocabulary.

Thus, the vocabulary reduction procedure will select the visual words that are more descriptive to certain actions. In this way, the selected visual words are expected to fulfil the following criteria:

- For a given action, the selected visual words should appear more frequently in this action. Also, they should be less common in videos that do not contain such an action.
- They should be frequently located on the moving person, even though they are surrounded by static objects or a cluttered background.

The two requisites conform to the TF-IDF weighting of information retrieval theory. These two clues are combined by the *VisualWordRank* algorithm [4], which leverages the idea of well-known *PageRank* [6]. In *PageRank*, a matrix is built to record the inherent importance of different webpages and the relationships among them. Iterations are carried out to update the weight of each webpage based on this matrix. After several iterations, the weights will stay stable and the final significance of each webpage is obtained combining both its inherent importance and the relationships with other webpages [6].

Based on this idea, we build matrix $R^{(C)}$ for each action category to combine the frequency and co-occurrence cues for visual vocabulary reduction. The elements in the diagonal of $R^{(C)}$ are defined as:

$$R_{i,i}^{(C)} = f_i^{(C)} / \ln(F_i) \tag{1}$$

where i is a VW and F_i and $f_i^{(C)}$ denote its average frequency in all categories and the within-category frequency in action category C, respectively. $R_{i,i}^{(C)}$ represents the inherent importance of visual word i. Thus, larger values of $R_{i,i}^{(C)}$ means that i is more important to category C.

The nondiagonal element $R_{i,j}^{(C)}$ is defined as the average co-occurrence frequency of visual words i and j in action category C. Therefore, if visual word i and j frequently co-occur in the category C, $R_{i,j}^{(C)}$ will present a high value. In order to detect the co-occurrence of word pairs, a spatial distance d is defined. A visual word pair co-occurrence is identified if the distance between these words is less than d. Because interest points may have various scales, the value of d is computed with

$$d = scale_i \times D \tag{2}$$

to achieve scale invariance, where $scale_i$ is the scale of the interest point from which the instance of visual word i is computed, and D a new parameter. In this work, experiments have shown that $D = 4$ is a good selection.

After computing $R^{(C)}$, its elements are normalized. Each visual word in the vocabulary is initially equally ranked and an iterative rank-updating is started. During the iterations, visual words having large inherent importance and strong co-occurrence (with large weights) will be highly ranked. After several iterations, the more relevant visual words can be identified by selecting the top N ranked or choosing the ones with rank values larger than a threshold. The detailed description of the *VisualWordRank* is presented in Algorithm 1 [4].

4 Experimental Results

In order to study the impact on accuracy of the vocabulary reduction in human action video classification, several tests have been performed using the KTH actions dataset [7]. This dataset consists of six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping, performed several times by 25 subjects in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors (s4). Fig. 3 shows frames examples of these videos with interest points extracted by STIP (yellow thick circles) and MoSIFT (green thin circles) superimposed. Note that the background is homogeneous and static in most sequences. The sequences were downsampled to the spatial resolution of 160×120 pixels and have a length of four

Algorithm 1 VisualWordRank [4]

Input: $R^{(C)}$; maximum iteration time: $maxiter$.

Output: The rank value of each VW to the category C:

$Rank_i{}^{(C)}, i = 1.....VW_{num}^{(C)}$

Initialize each element in the $VW_{num}^{(C)} \times 1$ sized rank vector $OldRank^{(C)}$ as 1.

Normalize the sum of each column of $R^{(C)}$ as 1.

Set $iter \leftarrow 0$.

While $iter < maxiter$

$NewRank^{(C)} \leftarrow R^{(C)} \cdot OldRank^{(C)}$

If $\left(\left|NewRank^{(C)} - OldRank^{(C)}\right| \leq \epsilon\right)$

break

End If

$OldRank^{(C)} \leftarrow NewRank^{(C)}$

$iter++$

End While

$Rank^{(C)} \leftarrow NewRank^{(C)}$

seconds on average. We follow the leave-on-out cross validation (LOOCV) evaluation method as it facilitates the performance comparison among different approaches.

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 [a] and MoSIFT [b] with the default parameters. The detected spatio-temporal features, about 500,000 and 1 million respectively, are first quantized into visual words and the video is then represented as the frequency histogram over the visual words. Vocabularies are constructed with *K-means* clustering. Eight independent executions of the *K-means* algorithm are performed and the run that finds out the best clustering is selected. To limit the complexity, we cluster a third of the training features randomly selected. Features are assigned to their closest vocabulary word using Euclidean distance. The resulting histograms of visual word occurrences are used as video sequence representations.

Vocabularies of different sizes (200, 600, 1000, 4000 and 10,000 visual words) are created using a different number of clusters for the *K-means* executions. These values are similar to several implementations found in the literature such as [2], [8], [9] and [10]. The vocabulary of 10,000 visual words is unusually large for this dataset. However, it is created in order to test the accuracy obtained when it is reduced to smaller sizes.

Several reduced instances of the initial vocabularies are generated by applying different thresholds to the maximum *VisualWordRank* value. In order to clearly illustrate the discriminative abilities of the reduced visual vocabularies, a simple classification based on the k nearest neighbor (KNN) and a histogram intersection metric is used. The training set has been employed to establish the most suitable value for k (k=5). Additionally, a Support Vector Machine (SVM), using the implementation in [11], has been utilized to classify actions to prove the entire performance of the vocabulary reduction algorithm in a real framework.

[a] www.di.ens.fr/ laptev/download.html

[b] lastlaugh.inf.cs.cmu.edu/libscom/downloads.htm

Tab. 1: Accuracies for the **STIP** descriptor obtained with different vocabulary sizes and their reductions for action classification with **KNN** in the KTH videos database. Best results are marked in bold.

#Words	Accuracy	#Words	Accuracy	#Words	Accuracy	#Words	Accuracy	#Words	Accuracy
200 (100%)	92.59	*600* (100%)	93.06	*1000* (100%)	92.59	*4000* (100%)	94.91	*10000* (100%)	94.44
188 (94%)	92.59	*560* (93%)	93.52	*838* (84%)	93.98	***3715* (93%)**	**94.91**	*9294* (93%)	94.44
***170* (85%)**	**93.98**	*456* (76%)	93.98	***785* (79%)**	**94.44**	*2794* (70%)	94.44	***7816* (78%)**	**94.44**
149 (75%)	93.52	***394* (66%)**	**94.44**	*676* (68%)	93.98	*2487* (62%)	94.44	*6308* (63%)	93.98
122 (61%)	89.35	*317* (53%)	92.59	*572* (57%)	90.28	*1631* (41%)	92.13	*4849* (48%)	93.06
99 (50%)	85.19	*274* (46%)	90.28			*1220* (31%)	90.74	*2912* (29%)	91.20
								1735 (17%)	88.43

Tab. 2: Accuracies for the **MoSIFT** descriptor obtained with different vocabulary sizes and their reductions for action classification with **KNN** in the KTH videos database. Best results are marked in bold.

#Words	Accuracy	#Words	Accuracy	#Words	Accuracy	#Words	Accuracy	#Words	Accuracy
200 (100%)	74.07	***600* (100%)**	**81.48**	*1000* (100%)	83.33	*4000* (100%)	87.04	***10000* (100%)**	**91.67**
182 (91%)	74.07	*548* (91%)	81.02	*831* (83%)	84.26	***3954* (99%)**	**89.35**	*9776* (98%)	91.20
***153* (77%)**	**76.39**	*354* (59%)	80.56	*638* (64%)	84.72	*3831* (96%)	87.96	*9506* (95%)	90.28
128 (64%)	73.61	*285* (48%)	80.56	***482* (48%)**	**85.19**	*3497* (87%)	87.96	*8775* (88%)	89.81
114 (57%)	73.15	*231* (39%)	78.70	*369* (37%)	83.80	*2798* (70%)	87.96	*4669* (47%)	88.43
99 (50%)	69.91			*275* (28%)	81.94	*1595* (40%)	87.04	*3897* (39%)	86.57
						918 (23%)	82.41	*1151* (12%)	80.56

Tab. 3: Accuracies for the **STIP** descriptor obtained with different vocabulary sizes and their reductions for action classification with **SVM** in the KTH videos database. Best results are marked in bold.

#Words	Accuracy	#Words	Accuracy	#Words	Accuracy	#Words	Accuracy	#Words	Accuracy
200 (100%)	94.31	*600* (100%)	94.67	*1000* (100%)	94.50	*4000* (100%)	96.00	*10000* (100%)	94.80
***188* (94%)**	**94.49**	*560* (93%)	94.67	***767* (77%)**	**94.50**	*3921* (98%)	96.00	*9051* (91%)	95.14
158 (79%)	94.32	***473* (79%)**	**94.67**	*729* (73%)	94.17	***3614* (90%)**	**96.00**	***7903* (79%)**	**95.30**
130 (65%)	92.47	*449* (75%)	93.50	*565* (57%)	94.17	*3230* (81%)	95.50	*6775* (68%)	94.46
112 (56%)	91.97	*389* (65%)	93.50	*501* (50%)	93.83	*2824* (71%)	95.00	*5785* (58%)	93.79
90 (45%)	89.30	*329* (55%)	91.17	*438* (44%)	91.33	*2156* (54%)	93.67	*4896* (49%)	91.44
						1632 (41%)	93.00	*2949* (29%)	90.42
						1090 (27%)	90.16	*1443* (14%)	85.28

Tab. 4: Accuracies for the **MoSIFT** descriptor obtained with different vocabulary sizes and their reductions for action classification with **SVM** in the KTH videos database. Best results are marked in bold.

#Words	Accuracy	#Words	Accuracy	#Words	Accuracy	#Words	Accuracy	#Words	Accuracy
200 (100%)	88.97	*600* (100%)	92.17	*1000* (100%)	93.67	*4000* (100%)	94.67	*10000* (100%)	93.99
***190* (95%)**	**88.97**	*578* (96%)	92.17	***922* (92%)**	**93.67**	*3922* (98%)	95.00	*8667* (87%)	94.33
182 (91%)	88.48	***540* (90%)**	**92.50**	*790* (79%)	93.00	***3855* (96%)**	**95.00**	***8013* (80%)**	**94.33**
136 (68%)	87.99	*494* (82%)	91.50	*441* (44%)	93.00	*3746* (94%)	94.67	*6698* (67%)	94.16
127 (64%)	86.81	*402* (67%)	91.50	*368* (37%)	91.17	*2771* (69%)	94.67	*5744* (57%)	93.49
105 (53%)	82.81	*308* (51%)	89.00	*304* (30%)	89.50	*1910* (48%)	94.00	*4290* (43%)	93.14
89 (45%)	78.63	*266* (44%)	88.50			*1086* (27%)	90.50	*2517* (25%)	90.64
								1523 (15%)	87.14

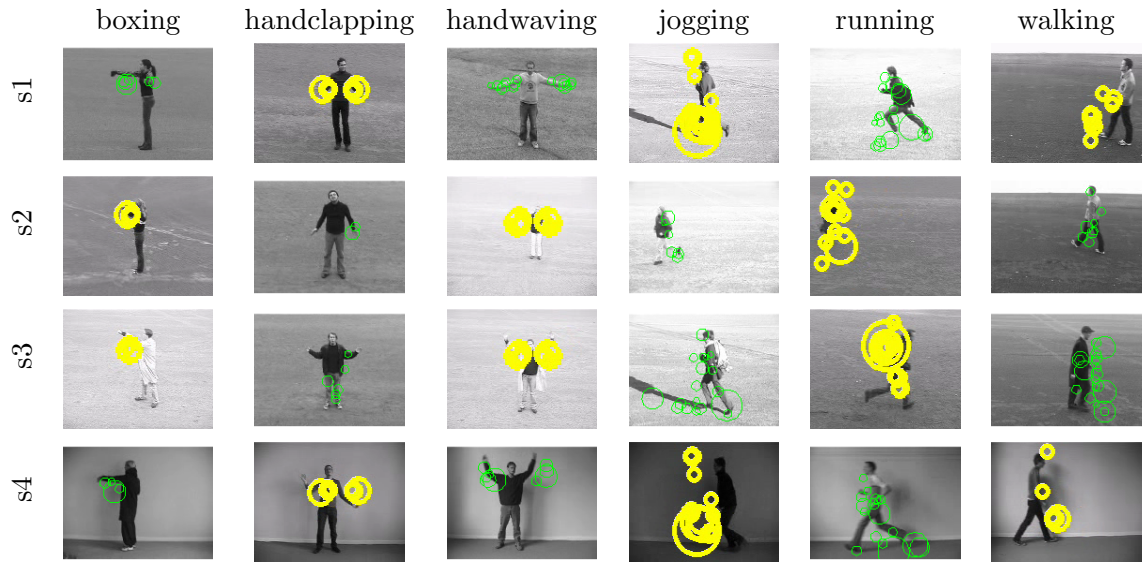

Fig. 3: Examples of interest points extracted by STIP (yellow thick circles) and MoSIFT (green thin circles) for different types of actions (columns) and scenarios (rows) of the KTH dataset.

Tables 1, 2, 3 and 4 summarize the most relevant results obtained. For each vocabulary (200, 600, 1000, 4000 and 10,000 words), the accuracy obtained with different size reductions and the percentage that these reductions represent with respect to the original vocabulary size are shown. As a general rule, accuracy is maintained and even increased with vocabulary size reductions in the range 20%-50%. However, this tendency is more noticeable with a KNN classifier than with the SVM classifier and with the STIP detector than with the MoSIFT detector. In the first situation, most people would agree that the SVM learning capabilities compensate the redundancies in the vocabularies. In the second one, it could be argued that the MoSIFT descriptor originates less redundant vocabularies. This general behaviour is particulary different in the case of a vocabulary size of 4000 words, where the best accuracies are reached and there is no room for much improvement. Additionally, more drastic vocabulary size reductions can be achieved if we can accept a slight loss in accuracy.

Finally, in order to demonstrate the validity of our method, Table 5 shows the performance reported by previous works using a similar experimental setup, i.e., classifying the whole videos rather that the different action sequences they contain and using the leave-on-out validation. As stated in [10], there are many variations in terms of experiment setups with other different methods, so a precise comparison is not possible. There are only two approaches outperforming the method proposed here. On one hand, Gao and Chen [10] use a parameter configuration for the extraction of MoSFIT descriptors specifically optimized for the KHT dataset that is not discussed in the paper. We believe that our results could outperform theirs by using their parameters for MoSIFT descriptors computation. By the other hand, Lui [12] solution needs additional training information manually annotated.

Tab. 5: Comparison with other methods all using the KTH dataset with a similar experimental set-up (LOOCV).

Method	Avg. Accuracy
Lui et al. 2010 [12]	97.00
Gao and Chen et al. 2010 [10]	96.33
Our method (STIP 3614 VWs)	**96.00**
Chen and Hauptmann 2009 [2]	95.83
Our method (MoSIFT, 3855 VWs)	**95.00**
Bergonzio et al. 2012 [13]	94.33
Liu and Shah 2008 [14]	94.20
Sun et al. 2009 [15]	94.00
Wong and Cipolla 2007 [16]	86.60
Niebles et al. 2008 [17]	81.50

5 Conclusions

The results presented in this paper proves that a reduction in the vocabulary size for human action description can improve the classification accuracy.

The experimental validation of the vocabulary reduction algorithm for action classification presented here is promising. However, an evaluation with more complex human action video databases is still necessary in order to obtain concluding results. We expect an improvement in the accuracy of the classifiers, as the complexity of these videos will originate richer vocabularies.

Additionally, further classification accuracy improvements are expected with the inclusion in the actions description of more spatial and temporal relationships among visual words.

References

[1] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[2] M.-Y. Chen and A. Hauptmann, "MoSIFT: Recognizing human actions in surveillance videos," Computer Science Department, Tech. Rep., 2009.

[3] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlatons," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2033 – 2040.

[4] S. Zhang, Q. Tian, G. Hua, Q. Huang, and W. Gao, "Generating descriptive visual words and visual phrases for large-scale image applications," *Image Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2664 –2677, 2011.

[5] D. Liu, G. Hua, P. A. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[6] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, no. 1-7, pp. 107–117, 1998.

[7] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *International Conference on Pattern Recognition (ICPR)*, vol. 3, 2004, pp. 32 – 36.

[8] Q. Hu, L. Qin, Q. Huang, S. Jiang, and Q. Tian, "Action recognition using spatial-temporal context," in *International Conference on Pattern Recognition (ICPR)*, 2010, pp. 1521 –1524.

[9] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, 2009, pp. 127–137.

[10] Z. Gao, M.-Y. Chen, A. G. Hauptmann, and A. Cai, "Comparing evaluation protocols on the KTH dataset," in *Proceedings of the First international conference on Human behavior understanding*, 2010, pp. 88–100.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[12] Y. M. Lui, J. R. Beveridge, and M. Kirby, "Action classification on product manifolds," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 833–839, 2010.

[13] T. H. Thi, L. Cheng, J. Zhang, L. Wang, and S. Satoh, "Structured learning of local features for human action classification and localization," *Image and Vision Computing*, vol. 30, no. 1, pp. 1 – 14, 2012.

[14] J. Liu and M. Shah, "Learning human actions via information maximization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[15] X. Sun, M. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 58–65.

[16] S.-F. Wong and R. Cipolla, "Extracting spatiotemporal interest points using global information," in *IEEE International Conference on Computer Vision (ICCV)*, oct. 2007, pp. 1 –8.

[17] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, vol. 79, no. 3, pp. 299–318, 2008.