

# Contextual Priming for Action Recognition in Realistic Videos

## Implementation and Evaluation of Baseline Approach

Sobhan Naderi Parizi  
Computer Vision and Active Perception Laboratory  
Royal Institute of Technology  
Stockholm, Sweden

November 26, 2009

### Abstract

In this report we will briefly go through the implementation details of a baseline approach towards realistic human action recognition. The experimental results will also be verified via comparison to the other evaluations which are done by other researchers on the same database. Finally, the results will be discussed and possible directions for further study will be suggested as well.

## 1 Introduction

In the previous report *Selection of database and feature descriptors* we motivated our interest in SIFT and HOF features for representing context and motion of human actions. Besides, we explored a wide range of different action databases and we well justified the reasons why we should work on HOHA2 database in the rest of our project. This is a challenging database recently introduced by Marszalek et al. Marszalek et al. (2009). In this report what we will present is basically an implementation of the method used in Marszalek et al. (2009). Unfortunately, the database contains two different training sets. The first set contains automatically labeled video samples<sup>1</sup> while the second set contains the manually labeled video samples. In other words, the first training set is noisy while the second one is clean. As we already mentioned, we will use the clean training set in order to be able to study the effect of different parameters on performance of the implemented method aside from the effect of labeling noise which might be unanticipated. Unfortunately, authors of Marszalek et al. (2009) has reported detailed experimental results only on the automatically annotated database. However, as we will see shortly in the next sections, we are lucky enough to be able to validate our results with a very recent paper Wang et al. (2009) which includes partial results on the same database.

In the next section we will briefly explain the implementation settings of the method and configuration of the utilized classifier. In Section 3 we will discuss our experimental results and also we will validate our achieved results by comparing them to the results reported by other researchers. Of course we have been careful to make sure the compared experiments are based on exactly the same database and the same settings. Finally, in Section 4 we will suggest the possible and promising directions for potential further studies.

## 2 Implementation Settings

The method is based on the Bag of Features (BoF) framework. In this method, first the features (SIFT and HOG+HOF features are used here) are extracted from the video (frame). Then a label is assigned to each feature vector based on index of its closest match to a visual vocabulary. The visual vocabulary is built from a randomly selected set of features of the same type using clustering algorithms. Since we have two different types of features here, i.e. SIFT and HOG+HOF, we will built two individual visual vocabularies, one for each type of feature. Similar to Marszalek et al. (2009) and Wang et al. (2009), we used 100,000 randomly selected features of each type and used K-Means to build our visual vocabularies. Moreover, to make the comparison fair, we also set the size of our visual vocabularies to be equal to 4000. However, unlike ?) we did not initialized our clusters 8 times. Nevertheless, we believe that this difference will not affect comparison of the two methods later on.

Importance of preserving spatial information in BoF framework is previously showed in several papers Lazebnik et al. (2006) Parizi et al. (2009). Therefore, similar to Marszalek et al. (2009), we will also use the same spatial and temporal grids on the extracted features.

Once the BoF histograms are built for all of the samples in the database, we use Support Vector Machines to classify the histograms based on one-against-all approach. In the next section we will discuss the classification mechanism in more details.

---

<sup>1</sup>The automatically labeled videos are weakly supervised in the sense that the labeling is done by textual analysis of the movie scripts which are publicly available for almost all of the movies nowadays.

**Table 1:** The table compares performance of three different feature descriptors, i.e. SIFT, HOG+HOF, and combination of the two, in terms of average precision (AP). Mean average precision (meanAP) for all of the 12 action classes is also shown at the bottom of the table.

Action Class	Average Precision (AP)		
	SIFT+hoghof	SIFT	hoghof
AnswerPhone	25.65	26.26	22.44
DriveCar	90.09	80.41	86.15
Eat	50.05	10.46	60.78
FightPerson	72.98	52.51	71.07
GetOurCar	43.78	36.72	36.29
HandShake	29.47	25.85	27.41
HugPerson	30.01	30.73	37.05
Kiss	59.32	49.75	53.00
Run	69.32	61.70	67.98
SitDown	55.25	44.95	55.07
SitUp	30.60	24.94	23.40
StandUp	62.28	50.42	55.37
<b>meanAP</b>	<b>51.57</b>	<b>42.23</b>	<b>49.67</b>

## 2.1 Classification of Feature Vectors using SVM

Recently, Support Vector Machine (SVM) classifiers have gained tremendous popularity in computer vision because of their excellent classification power in many problems. Here also SVMs are used for classification of the descriptor vectors. Following we briefly review mathematical formulation of the SVM classifiers.

In standard SVM classifier, the classification function can be formulated as follows:

$$f(X) = \sum \alpha_i y_i K(X^i, X) + b \quad (1)$$

where  $K$  is the kernel function and  $y_i$  is the class label associated with the  $i^{th}$  image. The  $b$  and  $\alpha_i$ s are the SVM parameters which will be optimized during training stage. We use RBF kernel with  $\chi^2$  distance metric which can be written as:

$$K_{RBF}(X^i, X^j) = \exp\left(\frac{-dist_{\chi^2}(X^i, X^j)}{\gamma}\right) \quad (2)$$

where  $\gamma$  is the variance measure of data points which can be optimized by doing grid search over a certain range of values. We can narrow down the search space for the optimal value of  $\gamma$  by using the estimate given by Equation 3. Therefore, we will limit our grid search in a certain vicinity of this estimated value.

$$\frac{1}{\gamma_{optimal}} \approx \frac{2}{N^2 - N} \sum_{i=1}^N \sum_{j=i+1}^N dist_{\chi^2}(X^i, X^j). \quad (3)$$

## 3 Experimental Results

In this section we will present performance of the implemented method in terms of average precision. For each action class, the AP value is reported individually. Moreover, we have evaluated 3 different feature types i.e. SIFT, hoghof, and combination of the two as illustrated in Table 1.

Although we have implemented the method used in Marszalek et al. (2009), but to the best of our knowledge, the evaluations that we have made here are not done by others. As we mentioned before, Marcin et al. have used the automatically labeled database in their evaluations. Yet another evaluations are made using the same method in Laptev et al. (2008), but, our database is not exactly the same database as theirs. However, fortunately, Wang et al Wang et al. (2009) has reported the mean average precision that they have achieved using hoghof features which is 47.60. As showed in Table 1, we estimated this value to be 49.67 in our evaluations. The two estimated performance values are quite close and the small difference could be because of the optimization of SVM parameters or the fact that we have not re-initialized our K-Means several times. Nevertheless, the difference is not so large that it disputes our evaluations.

### 3.1 Discussion

On of the conclusion made in Parizi et al. (2009) was that they showed that SIFT features basically captures more of the static appearance of the video scene while hoghof features captures dynamic information which is basically taken from the moving objects.

Looking back at Table 1, some interesting conclusions could be made about contribution of different features in recognition of certain classes of human actions which basically approves the conclusions made in Parizi et al. (2009). For example, in action classes such as *Run*, *Eat*, *FightPerson* that could happen in an extremely wide variety of scenes, SIFT features has relatively less contribution to the final classification compared to hoghof features. On the other hand for action classes such as *SitUp*, *GetOutCar* scene plays an integral role in recognition of the action. It is easy to think of the bedroom scene and car shape as the contextual structures that boost up performance of SIFT features in these action classes respectively. General performance of the three different features also is consistent with the experiments of Marszalek et al. (2009). Our results also proves that hoghof works better than SIFT in general and combination of the features will also increases the classification accuracy.

## 4 Future Works

Involvement of contextual information into human action recognition framework seems to be a hot topic. Among the recent works, Han et al. (2009) turns out to be the first paper which has used contextual information to improve the state-of-the-art method on realistic human action recognition scenarios. Their results also prove that context can be a very helpful type of information which have the potential to boost up action classification performance in the challenging databases such as HOHA1/2.

Regarding utilizing contextual information, there seem to exist a dichotomy of potential approaches which might yield an integral improvement on realistic action recognition problems. The first approach is to try to distinguish between the actor bounding box and the rest of the video frame and extract our features from each part individually. By doing this, we could get rid of the great deal of confusion that is made by the objects in the background which have no correlation with the action of interest (see Figure 1. The second approach is to brings temporal history of the features into account while capturing scene features of the frames meanwhile as well. We believe that temporal contextual information is also a big part of the information that people has been missing by using the traditional BoF framework so far.

**Figure 1:** Holistic BoF representations suffer from the confusion made by the other objects in the scene which are not of interest. Here, assuming that the goal is to recognize drinking action, the other guy which is smoking will perplex the BoF representation of the video. On the other hand, if we somehow could figure out that he is part of the background, this will help our representation of the drinking guy to be much more unique and isolated by focusing only on the drinking guy.



## References

- Dong Han, Liefeng Bo, and Cristian Sminchisescu. Selection and context for action recognition. 2009.
- Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. 2008.
- Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. 2006.
- Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. 2009.
- Sobhan Naderi Parizi, Ivan Laptev, and Alireza Tavakoli Targhi. Modeling image context using object centered grids. 2009.
- Heng Wang, Muhammad Muneeb Ullah, Alexander Klaaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. 2009.