Contextual Priming for Action Recognition in Realistic Videos

Sobhan Naderi Parizi
Under supervision of Prof. Danica Kragic
Computer Vision and Active Perception Laboratory
Royal Institute of Technology
Stockholm, Sweden

November 26, 2009

Abstract

In this report we will investigate the problem of contextual priming in realistic human actions. We consider the cases where the environment is uncontrolled. We will study different human action databases to find the one that best matches requirements of our problem. Furthermore, we will more or less decide what type(s) of feature descriptors would be interesting to be experimented. This decision is made based on an early anticipations of the experiments that we will need later in order to dissect influence of scene context in action recognition. Result of this report can hopefully be used in problems where the goal is to model the joint distribution of action and context. Furthermore, we will implement and briefly discuss the 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

During recent years much research has been done in the field of object detection and recognition in still images. Accordingly, very accurate and fast detectors have been established for faces (14), pedestrians (1), and many other object classes (3). Challenging databases and competitions (Everingham et al.) are organized due to the object detection problem in still images.

More recently people have moved from image framework to video framework. The results reported in (11) (2) (13) show impressing recognition performance of human actions in videos. More interestingly, the solutions are good enough to yield promising performance on real movies as well (5) (6) (10). Here the problem is to get a sequence of image frames, i.e. a video, and answer the question of which action class is performed within the video. Figure 1 shows some sample actions in real movies. As you see in the figure, the example could have a wide variety in terms of motion, background/forground appearance, viewpoint, etc.

Figure 1: Different action classes in realistic movies. Each row contains three samples of one of the action classes.

Some of the action classes may be context-free which means that inclusion of the scene in which the action is performed will bring no extra information about type of the action class. This could be due to two different reasons: The first possible reason is vast diversity of the context in which the action can be performed; for example sit-down action could be done within very many different scene contexts (6). Second reason could be the fact that for the context-free actions the video frame is almost completely filled by the actor and there is not much part of the scene context visible; for example the kissing action is most likely videotaped in a close frame view and it is expected to contain not much

part of the scene. Nonetheless, it is generally expected to have a strong correlation between type of the scene in which the action is performed and the action itself. For example sit-up action most likely happens in a bedroom scene context. Therefore, generally speaking, it is helpful to use scene features beside motion features in order to boost up accuracy of action recognition methods (10) (8). In this project we will basically try to incorporate scene features with motion features and study the mutual effect of these two types of features.

The rest of the report is organized as follows. Everything regarding selection of the database, plausible alternatives, as well as pros and cons of each of them is discussed in Section 2. In particular, in Section 2.2 we will specify the properties and configuration of the selected database in more details. Section 3 is devoted to the type(s) of feature descriptors that we will use. We will sketch a rough scheme of the experiments that we would do throughout the project based on which we justify selection of the feature descriptors. The rest of the report, basically, contains the implementation of the (10) paper. Particularly, in section 4.1 we go through the implementation details and Section 4.2 presents the achieved results. Finally, in Section 5 we will conclude the report.

2 Selection of Database and Feature Descriptors

As the first step into the project of *Contextual Priming for Action Recognition in Realistic Videos*, in this report we will investigate different human action databases and find the one that best matches requirements of the project. Furthermore, we will more or less decide what type(s) of feature descriptors to work with. This decision is made based on an early anticipations of the experiments that we will need later in order to dissect influence of scene context in action recognition. We would like to investigate the effect of motion/shape/color properties on the joint distribution of action and context.

The dataset that we are looking for should have special attributes including the followings:

i – We would like as much manual annotation as possible to be already available with the database

ii – Diverse and discriminant distribution of action classes a-priori conditioned over different scene classes

iii – Strong contribution of manipulated objects (not necessary for this project but can be quite useful for the sake of backward comparison in future e.g. when parameters of manipulated object are incorporated into the action model)

2.1 Which Database Shall We Choose?

Since our main goal is to study effect of scene context on action recognition problem, our primary requirement is that we need the actions to be scene dependent, as it naturally is in real world. For example, "PourCoffee" action is most likely performed in indoor scenes and most probably in kitchen. On the other hand "Run" action is most likely performed in outdoor scenes such as road or street. "Surgery", "HairMakeUp", and "Cooking" are more examples of the actions that have strong correlation with their scene context. From these examples it is obvious that we are interested in actions that are performed in their natural way without opposing strict limitations. In fact, what we are trying to model is the *natural* correlation that exists between an action and its corresponding scene. It implies that the environment should not be controlled. Therefore, most of the available databases are already out of the domain of our requirements such as the databases introduced in (16), (11), (13). Nonetheless, generally speaking, regardless of whether the action is constrained or not, we want to model the correlation between the actions and their performing scenes assuming that such a correlation exists. In this regard, we chose to work with real scenario actions since we believe that they do have the property.

There are three public datasets that more or less pass through most of our requirements which are introduced in (5), (6), (10). All of these datasets are dealing with action recognition from movies and therefore all of them are purely realistic. Furthermore, they are all annotated to some extent. Among these three, dataset (5) has an interesting property being the fact that the actor bounding box is annotated both in space and time which is exactly what we need. But, unfortunately, this database contains only two extremely similar actions, i.e. *Drink* & *Smoke*, which is not bad by itself. The fact that the two actions has an extremely large overlap in terms of their contextual information makes it unusable for our purpose. One more unpleasant issue about this dataset is that a part of the test database is videotaped by the authors and, therefore, is not realistic. They have recorded the samples in a fixed and constrained scene although several different actors were involved. Nonetheless, it could be very interesting if we get back to this dataset after this project when we want to incorporate manipulated object parameters into our action recognition model. More interestingly, the dataset contains bounding box annotation for the manipulated object as well; the annotation is only for the keyframe though being the frame in which the manipulated object reaches mouth of the actor.

As mentioned by the authors, there is no difference between the datasets of (6) and (10) except the fact that (10) is an extended version of (6). More interestingly, within the extended dataset they have included a separate scene database with corresponding annotation information. However neither the database of (6) nor (10) does not have annotation for actor bounding box which will be a bit of a problem for us.

2.2 HOHA2 Database

According to the aforementioned discussions, we choose to work with the dataset introduced in (10) which is named *Hollywood2* by the authors. It contains 12 action classes and 10 scene classes:

- {*AnswerPhone, DriveCar, Eat, FightPerson, GetOutCar, HandShake, HugPerson, Kiss, Run, SitUp, SitDown, StandUp*}
- {*Ext-house, Ext-road, Int-bedroom, Int-car, Int-hotel, Int-kitchen, Int-living-room, Int-office, Int-restaurant, Int-shop*}

The database contains several clips of video each of which contains *at least* one action. In a few cases there are two actions in one single video sample. Cases with three actions present in one clip also exist, although they are very rare. However, there is no sample clip which contains none of the 12 action classes at all. It means that one objection to this database, if used for action detection purpose, is that there is no background action in the database. By background action we mean cases where no recognized action is performed in the video.

The clips are are generated from 69 movies where the training samples are extracted from 33 and test samples are extracted from the rest 36 of them [1]. There are 3669 clips in total, covering both scene and action clips. There are two training sets in the database. One of them is manually annotated and the other one is automatically trained from the movie scripts and subtitles. Of course there is one manually annotated test set as well. In total, for all of the 12 action classes, there are 823 manually annotated training samples and 884 test samples. Since we want our evaluations to be as noise free as possible and we want to draw general conclusions based on comparison of our experiments, we would prefer to work with the manually annotated training samples.

For each of the 12 action classes there is an annotation file which includes a 1/-1 label corresponding to each of the samples (clips) determining whether the sample contains an action of that class or not. The same set of files exist for scene annotation.

For scene classes, besides the annotation files there are two probability distributions that are presented in two tables. The tables specify the conditional probability distributions $P(Scene|Action)$ and $P(Action|Scene)$. The authors have trained these probability tables based on statistical analysis of textual information that is available for the movies (such as subtitles and scripts). According to their findings, modeling the scene using this fixed probability tables leads to better action classification results compared to modeling the scene using visual features. Hopefully, we can show that our proposed approach, i.e. actor centered grid, can lead to a better visual modeling of scenes.

3 Selection of Feature Descriptor

Basically, we are interested in feature descriptors which have the following properties:

i – Describe motion/shape/color as good as possible

ii – Are invariant to scale/rotation

It has been shown in several papers that SIFT-like features are the best choice for scene classification (9)(7)(10). It is also intuitively obvious that distribution of color features is also a good indication of category of the scene. But, unfortunately, in the action database that we have chosen to work with, there are many black and white video samples. This forces us to only use simple SIFT features for scene classification. On the other hand, there may be interesting motion properties in the different scene classes conditioned on a specific action class. For example, for action classes such as *Run* an *DriveCar*, context of the video is most probably active and hence it can best be modeled by motion features; while, on the other hand, for *Eat* or *SitUP* actions, the context is presumably static i.e. their scene is better to be modeled in terms of appearance features rather than motion features. Therefore, it is assumed that the best feature descriptor for modeling scene of an action closely depends on type of the action itself. The alternatives are HOG and SIFT for appearance and HOF (Histogram of Optical Flow) and 3D-HOG for motion. According to the aforementioned remarks we choose SIFT over 2D-HOG. Between HOF and 3D-HOG, we choose HOF which has been shown to capture motion features better (10). In (5) the authors have shown that HOF works best on hard action recognition problem and its extension by concatenating with 3D-HOG will not help unless the two features are used as two completely different information channels.

As in (6) we use multi-scale approach to make the features robust to variation of scale.

4 Experimental Evaluations

In the previous section we motivated our interest in SIFT, HOG, 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. (10). In this report what we will present is basically an implementation of the method used in (10).

[1] Find full list of the movies at http://www.irisa.fr/vista/actions/hollywood2/

The database contains two different training sets. The first set contains automatically labeled video samples[2] 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 (10) 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 (15) 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 4.2 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 4.3 we will suggest the possible and promising directions for potential further studies.

4.1 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 (10) and (15), 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 (7) (12). Therefore, similar to (10), 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.

4.1.1 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:

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

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

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

where γ 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 γ by using the estimate given by Equation 3. Therefore, we will limit our grid search in a certain vicinity of this estimated value.

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

4.2 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 (10), 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 (6), but, our database is not exactly the same

[2]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
meanAP	**51.57**	**42.23**	**49.67**

database as theirs. However, fortunately, Wang et al (15) 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.

4.2.1 Discussion

On of the conclusion made in (12) 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 (12). 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 (10). Our results also proves that hoghof works better than SIFT in general and combination of the features will also increases the classification accuracy.

4.3 Future Works

Involvement of contextual information into human action recognition framework seems to be a hot topic. Among the recent works, (4) turns our 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 2. 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.

5 Conclusion

Throughout this report we investigated the problem of human action recognition in realistic scenarios and uncontrolled environment. First of all, this is a hard problem to be solved with the currently available methods. Unlike the impressive

Figure 2: 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.

results on many other datasets, such as KTH action database (13), the best performance reported on realistic action databases can hardly get to 50% in terms of mean average precision. Nevertheless, we believe that there are much untapped information that could boost up performance of the existing methods.

In this work, we implemented an existing action classification method and verified that one has to be careful in selecting the right feature descriptor in order to be able to capture the required information out of the video. We also verified that different features will contribute differently in representing different action classes. After all we proposed two different approaches that we believe are very likely to increase performance of the problem in hand in case they are further investigated.

References

[1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. 2005.

[2] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizig action at a distance. 2003.

[Everingham et al.] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[3] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. 2008.

[4] Dong Han, Liefeng Bo, and Cristian Sminchisescu. Selection and context for action recognition. 2009.

[5] Ivan Laptev and Patrik Perez. Retrieving actions in movies. 2007.

[6] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. 2008.

[7] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. 2006.

[8] Li-Jia Li and Li Fei-Fei. What, where and who? classifying events by scene and object recognition. 2007.

[9] Marcin Marszalek, Cordelia Schmid, Hedi Harzallah, and Joost van de Weijer. Learning representations for visual object class recognition. 2007.

[10] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. 2009.

[11] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. 2006.

[12] Sobhan Naderi Parizi, Ivan Laptev, and Alireza Tavakoli Targhi. Modeling image context using object centered grids. 2009.

[13] Christian Schuldt and Ivan Laptev. Recognizing human actions: A local svm approach. 2004.

[14] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. 2001.

[15] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaaser, Ivan Laptev, and Cordelia Schmid. Evaluaition of local spatio-temporal features for action recognition. 2009.

[16] Lihi Zelnik-Manor and Michal Irani. Statistical analysis of dynamic actions. *TPAMI*, 2006.