Contextual Priming for Action Recognition in Realistic Videos

Selection of database and feature descriptors

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

September 11, 2009

Abstract

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.

1 Introduction

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 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)

Everything regarding selection of the database, plausible alternatives, as well as good and bad points of each of them is discussed in Section 2.

After selection of the database, in Section 3, we have to decide on type(s) of the feature descriptors that we are going to use later throughout the project. If we already knew what are the set of experiments that we are going to do in the very last steps where we include the contextual modeling then we would have saved lots of effort by only focusing on those experiments when evaluating the baseline method. Since, at the moment, we still do not know what types of features will be used in the final setting of the project we then have to be very precautious in selecting the most efficient set of feature descriptors and avoid as much re-experiments as possible. Therefore, we will sketch a rough scheme of the experiments that we would do throughout the project and we will justify selection of the feature descriptors based on that.

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

i – Are state-of-the-art in human action recognition problem

ii – Are state-of-the-art in scene classification problem

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

iv – Are invariant to scale/rotation

2 Selection of Database

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 Zelnik-Manor and Irani (2006), Niebles et al. (2006), Schuldt and Laptev (2004). 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 Laptev and Perez (2007), Laptev et al. (2008), Marszalek et al. (2009). 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 Laptev and Perez (2007) 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 Laptev et al. (2008) and Marszalek et al. (2009) except the fact that Marszalek et al. (2009) is an extended version of Laptev et al. (2008). More interestingly, within the extended dataset they have included a separate scene database with corresponding annotation information. However neither the database of Laptev et al. (2008) nor Marszalek et al. (2009) does not have annotation for actor bounding box which will be a bit of a problem for us.

2.1 Specification of the Selected Database

According to the aforementioned discussions, we choose to work with the dataset introduced in Marszalek et al. (2009) 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

It has been shown in several papers that SIFT-like features are the best choice for scene classification Marszalek et al. (2007)Lazebnik et al. (2006)Marszalek et al. (2009). 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

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

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 Marszalek et al. (2009). In Laptev and Perez (2007) 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 Laptev et al. (2008) we use multi-scale approach to make the features robust to variation of scale.

3.1 Experiments

According to our current knowledge about the requirements of the project, it sounds that the following set of experiments would be enough to get the desired conclusions out of this project:

1. Baseline experiments (fixed spatial grid)
   (a) Modeling scene using SIFT BoF (Bag of Features)
   (b) Modeling scene using HOF BoF
   (c) Modeling action using HOF BoF
   (d) Merging the scene and action features together
2. Actor-centered experiments
   (a) Modeling actor-centered scene using SIFT/HOF
   (b) Modeling action using HOF BoF
   (c) Merging the scene and action features together (either in per-frame fashion or holistically)

It could also be a good experiment to model scene using 3D-HOG since it captures dynamic appearance (both motion and shape). But, we will delay this experiment until the end of the project and we will decide whether to do it or not based on the amount of time we had left at the end of the project.

We could also do temporal subdivision by making a grid along time. It has been shown in Laptev et al. (2008) that when we use spatio-temporal grids for making our final BoF histograms, type of the optimal grid, in terms of its geometric structure, is dependent on the class of action. Their final conclusion is that combination of different grid types works the best. They have shown that when we use only one type of grid, most often, the best grid type turns out to have no temporal subdivision. Nevertheless, when they combine different grid types, almost always, there exists a grid with multiple temporal subdivision in the combination of grid types that gives the best classification performance. We think that we may get different conclusions when we localize the grids based on the actor (actor-centered grids). Intuitively, temporal subdivision should be preferable assuming that we have perfect alignment both in terms of time (action frames) and image location (actor bounding box).

We think that it may be very much time consuming to experiment on different grid types with multiple subdivisions both along space and time because the length of the final feature vector is linearly increasing by the number of cells within the grid that we have chosen. It can end up in huge feature vectors and subsequently take a long time to be evaluated. Therefore, we will delay the decision about whether to use temporal subdivision or not until we get a sense of time complexity of the experiments.

References

Ivan Laptev and Patrik Perez. Retrieving actions in movies. 2007.

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, Cordelia Schmid, Hedi Harzallah, and Joost van de Weijer. Learning representations for visual object class recognition. 2007.

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

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

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

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