Correlations Between 48 Human Actions Improve Their Detection

G.J. Burghouts, K. Schutte
TNO Intelligent Imaging
gertjan.burghouts@tno.nl

Abstract

Many human actions are correlated, because of compound and/or sequential actions, and similarity. Indeed, human actions are highly correlated in human annotations of 48 actions in the 4,774 videos from visint.org. We exploit such correlations to improve the detection of these 48 human actions, ranging from simple actions such as walk to complex actions such as exchange. We apply a basic pipeline of STIP features, a Random Forest to quantize the features into histograms, and an SVM classifier. First, we show that the sampling for the Random Forest can be improved by exploiting the correlations between human actions. Second, we show that exploiting all 48 actions’ posteriors for detecting a particular action also improves further the detection in general. We demonstrate a 50% relative improvement for human action detection in 1,294 realistic test videos.

1. Introduction

We consider the challenge of automated detection of 48 human actions from 4,774 videos from the visint.org dataset. This dataset is novel (released early 2012), and it involves many complex actions. The actions vary from a single person (e.g. walk) to two or more persons (e.g. follow). There are actions that are defined by the involvement of some object (e.g. give), or an interaction with the environment (e.g. leave). The most complex actions involve two person and an object (e.g. exchange, throw-catch combination). This makes the problem in this paper very challenging, compared to the size and the complexity of other benchmarks, e.g. [1,2]. We argue that the complexity of simultaneously detecting 48 human actions, of which some are very complex, makes the faced problem very interesting.

From the state-of-the-art, we construct our baseline pipeline of STIP features [3], a Random Forest to quantize the features into histograms [4], and a SVM classifier with a $\chi^2$ kernel serving as a detector for each action. We aim to improve two important parts of this pipeline: (i) the selectivity of the front-end histograms, and (ii) the detection accuracy of the back-end classifier. We will do so by exploiting correlations between the 48 actions.

We start from the observation that particular human actions are very correlated, see Figure 1. These correlations have been computed from the annotations of the complete visint.org dataset, where each of the 4,774 videos has 48 present/absent annotations. One category of actions correlate by similarity, e.g. lift and raise. Another category are the sequential relations, e.g. hold and throw, as usually a person holds an object before throwing it. There is also a category of compound actions, e.g. to exchange something requires both to give and to receive something. A final example is the category of actions that involve two persons that need to be sufficiently close, like exchange. As people may initially be too far apart, such actions are correlated to walking, moving and approaching. Indeed, all of the actions from these examples, and many others, are highly correlated in the human annotations of the visint.org dataset.

The objective of this paper is to exploit such correlations to improve the detection of the 48 human actions. First, we will show that the sampling for the Random Forest can be greatly improved by exploiting the correlations between human actions. The correlations are helpful in selecting negative examples that humans consider similar to the positive examples. Indeed, all of the actions from these examples, and many others, are highly correlated in the human annotations of the visint.org dataset.

Second, we will consider the correlations again in a final detector for each action, by exploiting the posterior probabilities of all other actions. For correlated actions, their posteriors will to some degree be predictive of each other. We will demonstrate a relative improvement of 50% with respect to the baseline pipeline for human action detection in 1,294 realistic test videos, with highly varying recording conditions, with on average 195 test recordings of each action.

In Section 2, we describe the videos, annotations and experimental setup. Section 3 shows that the
selectivity of the Random Forest based human action histograms can be improved by selective sampling of its negatives. In Section 4, we consider all 48 human action detectors together, in order to exploit their correlations in a stacked classifier, and we demonstrate the improved detection rates for the 48 human actions. Section 5 summarizes the results.

Figure 1. Human actions in the visint.org dataset are highly correlated. Only correlations larger than 0.2 are shown. The thickest line indicates a correlation of 0.7.

2. Experimental Setup

Data. This experiment is about the classification of 48 human actions from 1,294 short test videos of 10-30 seconds, given a train set of 3,480 similar videos. This dataset is novel and contributed by DARPA on www.visint.org. The annotation is as follows: for each of the 48 human actions, a human has assessed whether the action is present in each video or not (“Is action X present?”).

Baseline pipeline. We consider the baseline pipeline as mentioned in the Introduction, where each Random Forests per action is grown, 10 trees and 32 leaves, based on 200K feature vectors, 100K from randomly selected positive videos, and 100K from randomly selected negative videos.

Advanced pipeline. The only parts we vary are the sampling of feature vectors for the Random Forest (Section 3) and the usage of posteriors for all 48 actions in a final-stage classifier (Section 4). All other parts are fixed during the experiment.

Performance measure. The performance will be measured by the Matthew’s Correlation Coefficient, or MCC measure, $MCC = (TP \cdot TN - FP \cdot FN) / \sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}$, where $T$=true, $F$=false, $P$=positive and $N$=negative. This performance measure is independent of the sizes of the positive and negative classes. This is important for our evaluation purpose, as there are +1,000 positive samples for “move”, to 61 samples for “bury”.

3. Selective Sampling of Negatives of Human Actions for Random Forest Generation

This section focuses on how to select the appropriate samples for creating a discriminative Random Forest. We do this by selective sampling. Such bootstrapping has been studied extensively [5]. Here we consider the problem of selecting the negative samples. For the positive class, of a particular human action, we hypothesize that we have a good sampling by random sampling. The total set of negative classes however, is much more heterogeneous: it can be any of the other 47 human actions. For discriminative learning, the rationale is that good samples are similar to the positive samples.

To select negatives that are similar to positives, social tags have been exploited in [6]. Those negatives are selected that have social tags similar but not identical to the target tag. Social tags are unbound as a user can input any textual string. That makes the search for good negatives hard and therefore an iterative scheme is applied to find a good subset of negatives.

In our case, the tags are bound: the annotations are fixed, as 48 binary judgements have been made by the annotators whether each action is present. We refer to the 48 annotations as the groundtruth vector. Due to the correlations between actions, we expect to be able to determine the similarity between videos based on their groundtruth vectors. For a particular human action, we have the groundtruths of the 47 other actions available that we will use to find negative samples that are similar to positive samples.

For each action, we sample randomly positive videos. We collect sufficient positive videos such that by extracting all feature vectors from each video we obtain 100K feature vectors. Next, we obtain another 100K feature vectors from negative videos. Together, these 200K feature vectors are used to create the Random Forest. This 100K + 100K setup bypasses the imbalance problem of having much more samples in the negative class [7]. The negative videos are obtained by either random sampling (Baseline Pipeline), or by our proposed scheme for selective sampling (Advanced Pipeline). We will compare the two variations.
Our selective sampling scheme searches for each positive video the negative video that has a groundtruth vector that is most similar. We compare the vectors with the L2 distance metric, where we normalize the vectors such that they sum to one. We continue to iterate over the set of positive videos, and including each iteration all feature vectors from the next most similar negative video, until we have collected the 100K feature vectors of the negative class.

Figure 2 shows the results of random vs. selective sampling. Both strategies result in a Random Forest, each with histograms for the train and test sets. From the train set histograms, we train an SVM with the $\chi^2$ kernel. We measure the detection accuracy. Selective sampling of negatives, yields an improvement for 21 out of 48 actions. For 13 actions the performance degrades. Relatively, the improvement is 12%.

4. Boosting Action Detection by Including other Actions’ Posteriors

In this section, our goal is to boost the performance further by using the posteriors from each action-specific SVM in a final-stage classifier. The rationale is that for correlated actions, their posteriors will, to some degree, be predictive of each other. For instance, the likelihood of walking is very predictive of moving.

We consider each action’s posterior to be a new feature value, leading to a new 48-dimensional feature vector, to be classified by a final-stage classifier. In previous work, different feature types, like image and text features in [8], were combined in a similar two-stage setup. In such approaches the features are heterogeneous and the target class is the same. In our case, the features are homogeneous (the input features are in all cases STIP, always represented by a Random Forest and classified by an SVM), whereas the target class is different, because we create a final-stage classifier for each of the 48 human actions. The challenge is how to combine the posteriors of each action in the final-stage classifier.

In [9], a weighted linear combination of each posteriors was proposed to yield a better estimate of each of the classes. We checked that a weighted linear combination is not a solution to our problem, as the human actions co-occur in particular configurations – and not generic, in the same configuration. For instance, “moving” may co-occur with walking, if it is a person who moves. If it is a car that moves, walking is unlikely. As a solution to this, we train an SVM classifier for each action specifically, based on the 48 posteriors.

The results are depicted in Figure 3. The final-stage combiner of the posteriors of all 48 human actions improves the detection for 32 out of 48 actions. On average, the MCC scores increase relatively with 50% compared to the first-stage action detectors.

For completeness, we compare these results to our early results [10] in DARPA’s Mind’s Eye program, where we achieved prominent performance during the MCC.
September 2011 evaluation trials. A significant gain of 118% is achieved.

We consider MCC scores > 0.3 a reasonable performance (for comparison, humans scored 0.5-0.7). We have achieved a reasonably good score for: Go, Walk, Leave, Dig, Run, Arrive, Pass, Flee, Bury, Lift, Jump, Pickup.

5. Conclusions

We have exploited correlations between 48 human actions to improve their detection. We use a pipeline of STIP features, a Random Forest to quantize the features into histograms, and an SVM classifier. First, we have shown that the sampling of train data for the Random Forest can be improved by exploiting the correlations between human actions. Selective sampling of videos, yields an improvement for 21 out of 48 actions. Relatively, the improvement is 12%. Second, we have demonstrated that exploiting all 48 actions’ posteriors for detecting a particular action further improves the detection in general. We have achieved an average relative improvement of 50%.

Acknowledgement

This work is supported by DARPA. The content of the information does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

References


Figure 3. The final-stage combiner of the posteriors of all 48 human actions improves with 50% on average compared to the first-stage action detectors.