# The LEAR submission at Thumos 2014

Dan Oneata, Jakob Verbeek, and Cordelia Schmid

Inria*

**Abstract.** We describe the submission of the INRIA LEAR team to the THUMOS workshop in conjunction with ECCV 2014. Our system is based on Fisher vector (FV) encoding of dense trajectory features (DTF), which we also used in our 2013 submission. This year's submission additionally incorporates static-image features (SIFT, Color, and CNN) and audio features (ASR and MFCC) for the classification task. For the detection task, we combine scores from the classification task with FV-DTF features extracted from video slices. We found that these additional visual and audio feature significantly improve the classification results. For localization we found that using the classification scores as a contextual feature besides local motion features leads to significant improvements.

## 1 Introduction

This paper describes our entry in the THUMOS Challenge 2014. The goal of the THUMOS Challenge is to evaluate action recognition approaches in realistic conditions. In particular the test data consists of untrimmed videos, where the action may be short compared to the video length, and multiple instances can be present in each video. For full details on the definition of the challenge, task, and datasets, we refer to the challenge website [3].

Below, we describe our systems for classification and detection in Section 2, and present experimental results in Section 3.

## 2 System description

We first describe our classification system to recognize untrimmed action videos in Section 2.1. The localization system presented in Section 2.2 is similar, but trained to recognize temporally cropped actions instead of complete untrimmed videos. The detection system also exploits the classification scores obtained for complete videos as a contextual feature.

### 2.1 Classification

For our classification system we build upon our winning entry in the THUMOS 2013 challenge. It is based on Fisher vector (FV) [8] encoding of improved dense trajectory features [9]. As last year we use a vocabulary of size 256, rescale the videos to be at most 320 pixels wide, and skip every second frame when decoding the video.

* LEAR team, Inria Grenoble Rhône-Alpes, Laboratoire Jean Kuntzmann, CNRS, Univ. Grenoble Alpes, France.

**Feature extraction.** This year, we have added several new features that complement the motion-based features. We add static visual appearance information through the following features:

1. **SIFT:** we extract SIFT features [6] on a dense multi-scale grid, and encode these in a FV using a vocabulary of size 1024. We extract SIFT on one frame out of 60, and aggregate all descriptors in a single FV.
2. **Color:** we extract color features based on local mean and variance of the color channels [1] every 60-th frame, and encode them in a single FV with a vocabulary size 1024.
3. **CNN:** we extract a 4K dimensional feature using a convolutional network trained on the ImageNet 2010 Challenge data. We use the CAFFE implementation [2], and retain the layer six activations after applying the linear rectification (which clips negative values to zero). We also experimented with using layer seven or eight, but found worse performance. We extract CNN features in every 10-th frame, and average them into a single video-wide feature vector.

In addition to the visual features, we also extract features from the audio stream:

1. **MFCC:** we down-sample the original audio track to 16 kHz with 16 bit resolution and then compute Mel-frequency cepstral coefficients (MFCC) with a window size of 25 ms and a step-size of 10 ms, keeping the first 12 coefficients of the final cosine transformation plus the energy of the signal. We enhance the MFCCs with their first and second order derivatives. The MFCC features are then aggregated into a FV with a vocabulary size of 256.
2. **ASR:** For ASR we used state-of-the art speech transcription systems available for 16 languages [4,5]. The files were processed by first performing speaker diarization, followed by language identification (LID) and then transcription. The system for identified language was used if the LID confidence score was above 0.7, else an English system as used. The vast majority of documents were in English, with a number in Spanish, German, Russian, French as well as a few in 8 other languages. Therefore, we only used the English transcripts, and represent them using a bag-of-word encoding of 110K words.

**Classifier training.** To train the action classification models, we train SVM classifiers in a 1-vs-rest approach. We perform early fusion to the dense trajectory features, by concatenating FVs for the MHB, HOG, and HOF channels. Similarly we early fuse the two local image features: SIFT and color. We, then, learn a per-class late-fusion of the SVM classifiers trained on the early fusion channels and the CNN, MFCC, and ASR features.

We also investigated the effect of using different parts of the training data. The *Train* part consists of 13,320 trimmed action clips across the 101 action classes. The *Validation* part consists of 1,010 untrimmed videos across the 101 action classes (10 per class), which are representative for the test videos. Finally, the *Background* part consists of 2,500 untrimmed videos not corresponding to any of the action classes.

### 2.2 Localization

To assess our performance we split the 1010 videos from the *Validation* split into two equal parts; we used one of them as train split and the other one as test.

For the temporal action localization task we only use the dense trajectory features, since the remaining features are more likely to capture contextual information rather than information that can be used for precise action localization.

We train 1-vs-rest SVM classifiers, albeit using only trimmed action examples from the *Train* and *Validation* sets as positives. As negatives we use (i) all examples from other classes of the *Train* part of the data, (ii) all untrimmed videos in the *Background* part of the data, (iii) all untrimmed videos of other classes in the *Validation* part of the data, and (iv) all trimmed examples of other classes in the *Validation* part of the data. In addition we performed one round of hard-negative mining on the *Validation* set, based on a preliminary version of the detector, and used these as additional negatives.

For testing we use temporal detection windows with a duration of 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, and 150 frames, which we slide with a stride of 10 frames over the video. After scoring the windows, we apply non-maximum suppression to enforce that non of the retained windows are overlapping.

Following [7], we re-score the detection windows by multiplying the detection score by the duration of the window. This avoids a bias towards detecting too small video fragments. In addition, we experimented with a class-specific duration prior, estimated from the training data.

Finally, we combine the window's detection score with the video's classification score for the same action class. This pulls-in additional contextual information from the complete video that is not available in the temporal window features. We take a weighted average of these scores; the weight is determined using the *Validation* set.

## 3 Results

In this section we present experimental results obtained on the *Validation* set.

### 3.1 Classification results

For the classification task we split the *Validation* set into 30 train/test folds. For each training fold we select 7 samples from each class, with the test fold containing the remaining 3 samples. We report the mean and the standard deviation of the mAP score across these 30 folds.

Table 1 presents an evaluation of the invidual features. The results show that the visual features are the strongest, in particular the motion features. Combining features significantly improves the results, *e.g.* from 52.02% mAP for MBH, to 64.35% for MBH + HOF + HOG. When combining all features, we obtain 77.84% mAP. Interestingly, the high-level ASR feature brings more than 4% mAP improvement when all other features are already included.

Next, we evaluate the effect of using different parts of the training data and test on the held-out part of the validation set, see above descprition of the cross-validation

(a)

| Feature | mAP |
|---|---|
| MBH | 52.02 ± 2.4 |
| HOF | 50.38 ± 1.9 |
| HOG | 48.79 ± 2.3 |
| CNN | 48.42 ± 2.0 |
| Color | 37.36 ± 1.7 |
| SIFT | 37.17 ± 1.8 |
| ASR | 20.77 ± 1.0 |
| MFCC | 18.97 ± 1.5 |

(b)

| Early fusion | mAP |
|---|---|
| EF1: MBH + HOF + HOG | 64.35 ± 2.3 |
| EF2: SIFT + Color | 45.78 ± 2.3 |
| Late fusion | |
| LF1: EF1 + EF2 | 69.62 ± 2.18 |
| LF2: EF1 + EF2 + CNN | 71.06 ± 2.00 |
| LF3: EF1 + EF2 + CNN + MFCC | 73.65 ± 1.90 |
| LF4: EF1 + EF2 + CNN + ASR | 76.26 ± 1.85 |
| LF5: EF1 + EF2 + CNN + MFCC + ASR | 77.84 ± 1.70 |

**Table 1.** Evaluation of individual features (a) and combinations (b) for the classification task.

| Validation | Y | | Y | Y | | Y |
|---|---|---|---|---|---|---|
| Train | | Y | Y | | Y | Y |
| Background | | | | Y | Y | Y |
| LF5 mAP | 70.40± 1.6 | 68.74 ± 2.2 | 77.84 ± 1.7 | 67.94 ± 1.9 | 67.90 ± 2.2 | 77.70 ± 1.8 |

**Table 2.** Evaluation of different parts of the training data for the classification task.

procedure. The results in Table 2 clearly show the importance of using both the trimmed (in *Train*) and untrimmed (in *Validation*) examples; untrimmed videos are important since these are representative of the test set, and the trimmed examples are important because they are roughly 10 times more of them. The videos in the *Background* set were not useful, probably because there are enough negative samples across the *Train* and *Validation* dataset. In conclusion, we used the *Train* and full *Validation* sets in our submitted classification results.

### 3.2 Localization results

For our localization system we have to compute features and scores for many temporal windows, and this is much more costly than the classification of entire videos. Therefore, we first evaluated the effect of using only MBH or all three trajectory features, and the impact of using a smaller vocabulary of size 64 *vs*. using the one of size 256 used for classification. In these experiments we follow [7], and rescore the windows using their duration. The first three rows of Table 3 show that the performance drops significantly if we use a smaller vocabulary, or use only MBH features. Therefore, we keep all trajectory features and the vocabulary of size 256 in all remaining experiments.

In the remaining experiments in Table 3 we consider the benefit of including the classification score as a contextual feature to improve the localization performance. The trade-off between the classification and detection score is determined cross-validation. The classification and detection scores are first normalized to be zero-mean and unit-

| System | Rescoring | Remarks | mAP |
|---|---|---|---|
| D1 | clip duration | K=64, MBH | 12.56 |
| D2 | clip duration | K=64, MBH + HOF + HOG | 14.58 |
| D3 | clip duration | K=256, MBH + HOF + HOG | 19.17 |
| D3+C, $\lambda = 0.2$ | clip duration | Run #3 | 21.63 |
| D3+C, $\lambda = 0.2$ | class specific prior, *Train+Val.* | | 21.57 |
| D3+C, $\lambda = 0.25$ | class specific prior, *Validation* | Run #1 | 26.57 |
| D3+C$^*$, $\lambda = 0.25$ | class specific prior, *Validation* | Run #2, C$^*$ visual-only | 26.52 |
| D3 | class specific prior, *Validation* | | 24.43 |

**Table 3.** Evaluation of action localization using the detection (D) and classification (C) system. The combined score is a weighted average which weights the detection score by $\lambda$ and the classification score by $(1 - \lambda)$.

variance so that the scores are comparable, and the combination weight has a natural interpretation. In the first experiment (row 4) we combine the best detector D3 (with mAP 19.17%) with the classification model using all our channels, which leads to an improved mAP of 21.63%. This is the system submitted as Run #3.

Instead of rescoring with the clip duration, we also considered rescoring with a class-specific prior on the duration (obtained using a histogram estimate). This leads to a similar performance of 21.57% mAP.

We observed a difference in the duration distribution of positive action instances in the *Train* and *Validation* part of the data, see Figure 1. This difference is explained by different annotation protocols and teams used to annotate these parts of the data. Therefore, we also considered using a prior estimate based on the validation data only. This leads to a significantly improved localization mAP of 26.57%. This is the system we submitted as Run #1.

Finally, submitted Run #2 is similar to Run #1, but is a vision-only run that excludes the MFCC and ASR audio features in the classification model. The system corresponding to the Run #2 obtains a performance of 26.52% mAP on our test split. Interestingly, in this case the audio features do not have a signifiant impact. To verify that the detection still benefits from the classifier when using the stronger prior, we also include a last run that uses this prior without the classification score (last row). This leads to a reduction in performance to 24.43%, showing that global video context is useful in the localization task, even when using the strong prior on duration.

## 4 Conclusion

In this notebook paper we have described our submission to the THUMOS 2014 Challenge, and presented an experimental evaluation of its components. Our main findings are as follows. (i) Additional visual and audio features significantly improve over a system based on dense trajectory features only (as we used in our winning entry in the

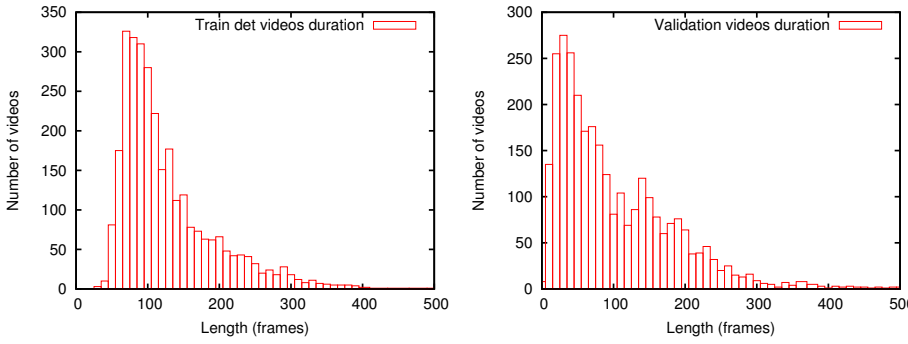


**Fig. 1.** Duration histograms of positive action instances across the 20 classes used for localization for the *Train* (left) and *Validation* (right) part of the data.

2013 THUMOS Challenge). This improved our results from 64.35% mAP to 77.84% mAP in our evaluation. (ii) For action classification in untrimmed videos it is beneficial to include representative untrimmed training videos in addition to trimmed action examples. This improved our results from 68.74% mAP to 77.84% mAP in our classification experiments. (iii) For action localization in untrimmed videos it is beneficial to use global video features, which we included in the form of the video classification scores. This improved our results from 19.17% mAP to 21.63% mAP in our localization experiments. (iv) For action localization it is important to include a rescoring based on the clip duration, a class specific prior estimated from the validation data worked best and improved our results from 21.63% mAP to our best result of 26.57%.

## Acknowledgements

We would like to express our gratitude to Lori Lamel and Jean-Luc Gauvain of the CNRS LIMSI laboratory[1] for providing the ASR transcripts. The ASR systems were partially developed within the Quaero program.[2] We also would like to thank the Fraunhofer Institute[3], our partner in the EU project AXES[4], for providing the MFCC code. This work was partly supported by the European integrated project AXES and the ERC advanced grant ALLEGRO.

## References

1. Clinchant, S., Renders, J.M., Csurka, G.: Trans-media pseudo-relevance feedback methods in multimedia retrieval. In: Advances in Multilingual and Multimodal Information Retrieval (2008)

[1] http://www.limsi.fr/tlp
[2] http://www.quaero.org
[3] http://mmprec.iais.fraunhofer.de
[4] http://www.axes-project.eu

2. Jia, Y.: Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org (2013)
3. Jiang, Y.G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14 (2014)
4. Lamel, L.: Multilingual speech processing activities in Quaero: Application to multimedia search in unstructured data. In: Human Language Technologies - The Baltic Perspective (2012)
5. Lamel, L., Gauvain, J.L.: Speech processing for audio indexing. In: Advances in Natural Language Processing (2008)
6. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
7. Oneata, D., Verbeek, J., Schmid, C.: Action and event recognition with Fisher vectors on a compact feature set. In: ICCV (2013)
8. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the Fisher vector: Theory and practice. IJCV 105(3), 222–245 (2013)
9. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV (2013)